А.Н. Кричевец, Е.В. Шикин,
А.Г. Дьячков

МАТЕМАТИКА ДЛЯ ПСИХОЛОГОВ

УЧЕБНИК

ФЛИНТА

МПСИ

А.Н. Кричевец,
Е.В. Шикин, А.Г. Дьячков

МАТЕМАТИКА ДЛЯ ПСИХОЛОГОВ

Учебник

Рекомендовано
Редакционно-издательским Советом
Российской Академии образования
к использованию в качестве
учебно-методического пособия

Издательство «Флинта»
Московский психолого-социальный институт
Москва
2003

УДК 22.1
ББК 51-7
К82

Кричевец А.Н.

Математика для психологов: Учебник / А.Н. Кричевец, Е.В. Шикин, А.Г. Дьячков / Под ред. А.Н. Кричевца. — М.: Флинта: Московский психолого-социальный институт, 2003. — 376 с.

ISBN 5-89349-400-8 (Флинта)
ISBN 5-89502-492-0 (Московский психолого-социальный институт)

Цель пособия — разъяснение основных математических понятий, необходимых в работе психолога. Книга состоит из четырех разделов: линейная алгебра и аналитическая геометрия; математический анализ; теория вероятностей и основы математической статистики. Изложение материала ориентировано на последующие приложения в психологии и сопровождается примерами из современной психологической литературы.

Для студентов психологических факультетов вузов.

ISBN 5-89349-400-8 (Флинта)
ISBN 5-89502-492-0 (Московский психолого-социальный институт)

Оглавление

Часть III. Теория вероятностей 203

Предисловие

Психология — наука многоплановая. Психологами в равной степени мы называем психолога-исследователя, использующего новейший томограф для описания мозговой активности и записывающего и обрабатывающего мегабайты цифровой информации, психолога-психотерапевта, беседующего с лежащим на психоаналитической кушетке клиентом, нейропсихолога, пытающегося найти причины плохого почерка у данного школьника и предложить способы преодоления возникающих в связи с этим трудностей обучения, — и многих других профессионалов, которые не всегда даже понимают друг друга.

При таком многообразии специальностей трудно надеяться, что единый курс математики может обеспечить нужды всех и каждого, не нагружая при этом учащегося материалом, который никогда не пригодится будущему специалисту. Дело еще более осложняется, если учесть, что курс математики приходится на начальные годы подготовки психолога, когда рядовой студент еще слабо представляет, куда в результате заведет его нелегкое вузовское поприще и какие отрасли математики окажутся необходимы ему уже в ближайшем будущем.

Именно на эту ситуацию *неопределенности в квадрате*, как сказал бы математик, и рассчитан данный учебник. Он состоит из двух “слоев”.

Первый, доступный в полной мере любому выпускнику российской школы, представляет собой почти минимальную базу для усвоения материала всех основных курсов по специальности “психология”. Изложение здесь в основном ведется на примерах, утверждения чаще не доказываются, а объясняются. Материал компонуется таким образом, что читатель получает что-то, даже если читает только первые несколько глав, поскольку многие важные математические идеи формулируются и объясняются почти в самом начале разделов и лишь затем уточняются и вписываются в более широкий математический контекст. Этот материал сосредоточен в главах с нечетными номерами.

Параллельный “слой” четных глав построен в традиционной манере математических учебников, включает формулы, теоремы и доказательства. Иногда материал нечетной главы повторяется в более строгом

виде в соответствующей четной, иногда там дается дополнительная информация, иногда — требующие математических выкладок разъяснения.

Каковы возможные стратегии работы с книгой?

Безусловно, возможно чтение только нечетных глав, дающее связное представление об основных математических понятиях и формирующее основные умения и навыки. Также, безусловно, возможно чтение книги подряд, не обращая внимания на некоторые повторы. В первых двух частях, посвященных линейной алгебре и математическому анализу, способный читатель может ограничиться только четными главами, но что касается теории вероятностей и математической статистики, то здесь продвинутому читателю рекомендуется чтение всего материала.

Подразумевается также возможность последующей работы с четными главами на старших курсах, когда сфера интересов студента становится более определенной. В этом случае задача облегчается тем, что продвижение будет вестись в виде систематического уточнения уже усвоенного материала нечетных глав.

В книге 1-я и 2-я части написаны А.Н. Кричевцом, 3-я часть — Е.В. Шикиным, А.Н. Кричевцом, 4-я часть — А.Г. Дьячковым, А.Н. Кричевцом, Е.В. Шикиным.

Часть I

Линейная алгебра

Глава 1

Линейные уравнения (идеи и примеры)

В этой главе мы будем рассматривать уравнения, причем такие, которые выглядят наиболее просто. Они содержат неизвестные только в первой степени и называются линейными. Линейным является следующее уравнение:

$$5x + 17 = 0.$$

Также линейны уравнения

$$y = 1 \qquad \text{и} \qquad 123456y + 654321y + 9z = 0.$$

Не являются линейными уравнения

$$x^2 + y = 1 \qquad \text{и} \qquad y^x + 1 = 17.$$

Решить уравнение — это значит найти такое число, которое превращает уравнение в равенство. Например, число 17 является решением уравнения

$$2x = 34,$$

а пара чисел $(1; 2)$ являются решением уравнения

$$x + y = 3$$

(если 1 подставить вместо x, а 2 вместо y). Заметим, что пара (2,1; 0,9) также является решением этого уравнения, как и бесконечно много других пар чисел.

Для обозначения неизвестных на первое время нам хватит букв: x, y, z, хотя несколько позже нам придется ввести новые обозначения.

Нас будут интересовать системы линейных уравнений — например, такие:

$$\begin{cases} 3x + y + 2z = 5 \\ 2x + y + z = 4 \\ x + 2y + z = 5. \end{cases}$$

Фигурная скобка перед набором уравнений указывает на то, что они представляют собой систему. В данном случае тройка чисел $(1;2;0)$ является решением каждого линейного уравнения данной системы, что и означает по определению, что это решение системы в целом.

1.1. Метод Гаусса

Существует множество различных методов решения систем линейных уравнений. Мы изложим один из них — метод Гаусса. Отметим, что умение решать уравнения в данном случае совсем не главное из того, что мы надеемся приобрести, освоив данный метод. Итак, решаем методом Гаусса следующую систему линейных уравнений

$$\begin{cases} 2x + 4y + 2z = 4 \\ 5x + 2y + 2z = 5 \\ 3x + y + z = 3. \end{cases} \tag{1.1}$$

Сначала разделим обе части первого уравнения на 2 и запишем новое уравнение на место прежнего.

$$\begin{cases} x + 2y + z = 2 \\ 5x + 2y + 2z = 5 \\ 3x + y + z = 3, \end{cases}$$

теперь умножим его на 5 и запишем отдельно

$$5x + 10y + 5z = 10.$$

С помощью этого варианта первого уравнения преобразуем второе уравнение системы, вычитая из левой части второго уравнения левую часть первого, а из правой — правую и приводя затем подобные члены. Получим новый вариант второго уравнения

$$- 8y - 3z = -5,$$

который запишем на второе место:

$$\begin{cases} x + 2y + z = 2 \\ \quad - 8y - 3z = -5 \\ 3x + y + z = 3. \end{cases}$$

Мы выполнили один шаг метода Гаусса и избавились от переменной x во втором уравнении. На втором шаге произведем аналогичную операцию с первым и третьим уравнениями, чтобы избавиться от переменной x в третьем уравнении. Для этого умножим первое на 3 и вычтем результат из второго. То, что получится, запишем на третье место:

$$\begin{cases} x + 2y + z = 2 \\ \quad - 8y - 3z = -5 \\ \quad - 5y - 2z = -3. \end{cases}$$

Теперь последний шаг в данном примере. Разделим новое второе уравнение на -8 и поменяем на противоположные знаки в третьем уравнении системы:

$$\begin{cases} x + 2y + z = 2 \\ \quad y + \frac{3}{8}z = \frac{5}{8} \\ \quad 5y + 2z = 3. \end{cases}$$

Умножим новое второе уравнение на 5 и вычтем результат из нового третьего уравнения, чтобы избавиться в нем от переменной y:

$$\begin{cases} x + 2y + z = 2 \\ \quad y + \frac{3}{8}z = \frac{5}{8} \\ \quad \frac{1}{8}z = -\frac{1}{8}. \end{cases} \tag{1.2}$$

Система приведена к «треугольному» виду и тем самым фактически решена. Последнее уравнение дает $z = -1$. Подставляя во второе, получаем $y = 1$. Подставляя оба значения в первое уравнение, получаем $x = 1$. Метод Гаусса приводит систему к виду, когда ее решение очевидно. Не всегда это — «треугольный» вид, о чем свидетельствует следующий пример.

$$\begin{cases} x - y - z = 1 \\ x + y + z = 3 \\ \quad y + z = 1. \end{cases} \tag{1.3}$$

После вычитания первого уравнения из второго получаем

$$\begin{cases} x - y - z = 1 \\ \quad 2y + 2z = 2 \\ \quad y + z = 1. \end{cases}$$

Поскольку второе и третье уравнения пропорциональны, следующий шаг метода Гаусса приводит к результату:

$$\begin{cases} x - y - z = 1 \\ \quad y + z = 1 \\ \quad 0 = 0. \end{cases}$$

У этой системы много решений. Это, например, тройки чисел $(2;1;0)$ и $(2;2;-1)$.

Упражнение 1.1. Описать все решения данной системы.

Совсем немного изменив последнюю систему, мы можем получить кардинально иной результат.

$$\begin{cases} x - y - z = 1 \\ x + y + z = 3 \\ \quad y + z = 2. \end{cases}$$

После первого преобразования получаем

$$\begin{cases} x - y - z = 1 \\ \quad 2y + 2z = 2 \\ \quad y + z = 2. \end{cases}$$

Продолжая действовать методом Гаусса, на месте третьего уравнения получаем странное равенство:

$$\begin{cases} x - y - z = 1 \\ \quad y + z = 1 \\ \quad 0 = 1. \end{cases}$$

Система, преобразования которой приводят к такому результату, называется *несовместной*, она не имеет ни одного решения.

В общих чертах изложение идеи метода Гаусса закончено. Однако некоторые его моменты мы умышленно оставили в тени, и теперь пришло время их обсудить.

Рассмотрим два вопроса.

1) Что делать, если после первых двух шагов метода получилась, например, такая ситуация:

$$\begin{cases} x - 2y - 3z = 4 \\ 2z = 2 \\ z = 1. \end{cases}$$

2) Применим ли метод Гаусса к такой системе:

$$\begin{cases} 0x + y - z = 1 \\ 0x + 2y + 2z = 2 \\ 0x + 3y + z = 3. \end{cases}$$

Эта система трех уравнений с тремя неизвестными имеет "нулевой столбец" переменная x входит в систему только с нулевыми коэффициентами. Нужно ли вообще рассматривать таких "уродов"? Математики уверенно отвечают утвердительно на последний вопрос, даже если они не могут сразу указать ситуацию, в которой такая система может появиться (у нас такие системы появятся довольно скоро и вполне органично).

В приведенных примерах в обоих случаях следует поменять местами столбцы. В первом случае получаем

$$\begin{cases} x - 3z - 2y = 4 \\ 2z = 2 \\ z = 1. \end{cases}$$

Дальше вычитание второго уравнения, деленного на 2, из третьего приводит к уже знакомой ситуации.

Во втором случае, переставив столбцы, получим

$$\begin{cases} y - z + 0x = 1 \\ 2y + 2z + 0x = 2 \\ 3y + z + 0x = 3. \end{cases}$$

После уже знакомых нам преобразований получаем

$$\begin{cases} y - z + 0x = 1 \\ z + 0x = 0 \\ 0 = 0. \end{cases}$$

Решением системы будет следующая комбинация: x — любое число, $y = 1, z = 0$. Заметим еще, что в случае получения нулевого уравнения может понадобиться также и перестановка строк системы.

Результат применения метода Гаусса к произвольной системе линейных уравнений — это система “ступенчатого” вида.

В системе ступенчатого вида

1) первый коэффициент первого уравнения не равен 0 (если в системе есть хотя бы один отличный от нуля коэффициент — эту оговорку мы поясним в следующей главе);

2) если в системе больше одного уравнения, то первый коэффициент второго уранения равен 0, а второй его коэффициент не равен 0;

. . . .

n) если в системе больше чем $n-1$ уравнений, то первые $n-1$ коэффициентов n-го уравнения равны 0, а n-й коэффициент не равен 0.

Упражнение 1.2. Убедиться, что система

$$\begin{cases} y - z + 0x = 1 \\ x = 0 \\ 0 = 0 \end{cases}$$

не подходит под наше определение системы ступенчатого вида.

1.2. Однородные системы линейных уравнений

Однородной системой линейных уравнений называется система, в которой в правой части каждого уравнения стоит 0. Вот пример такой системы:

$$\begin{cases} 2x + 4y + 2z = 0 \\ 5x + 2y + 2z = 0 \\ 3x + y + z = 0. \end{cases}$$

Ее мы получили из ранее решенной системы, которая на полях была помечена знаком (1.1), подставив в правую часть каждого уравнения 0. Понятно, что, применяя метод Гаусса к этой системе, мы получим предсказуемый результат, отличающийся от полученного ранее (1.2) только нулями в правой части.

$$\begin{cases} x + 2y + z = 0 \\ y + \frac{3}{8}z = 0 \\ \frac{1}{8}z = 0. \end{cases}$$

Точно так же однозначно получается и решение системы: из третьего уравнения $z = 0$, подставляя далее, получим $y = 0; x = 0$. Можно было бы сказать заранее, что система имеет нулевое решение, однако без

приведения к треугольному виду нельзя гарантировать, что оно единственное. Рассмотрим следующий пример, который получен из (1.3), заменой правых частей на нули.

$$\begin{cases} x - y - z = 0 \\ x + y + z = 0 \\ \phantom{x + {}} y + z = 0. \end{cases}$$

После преобразований, аналогичных проделанным с системой (1.3), получаем ступенчатую систему

$$\begin{cases} x - y - z = 0 \\ \phantom{x - {}} 2y + 2z = 0 \\ \phantom{x - 2y + {}} 0 = 0. \end{cases}$$

Мы получили фактически два уравнения с тремя неизвестными. Преобразуем систему в такой вид:

$$\begin{cases} x - y = z \\ \phantom{x - {}} 2y = -2z. \end{cases}$$

При любом значении, которое мы можем присвоить переменной z, мы получим треугольную систему двух уравнений с двумя неизвестными, которая имеет единственное решение. Действительно, пусть $z = 17$. Тогда система примет вид

$$\begin{cases} x - y = 17 \\ \phantom{x - {}} 2y = -34, \end{cases}$$

откуда получаем $y = -17$; $x = 0$.

Будем говорить, что мы имеем одну степень свободы при выборе решения. Одна степень свободы — это произвольный выбор одного параметра, по которому остальные переменные определяются однозначно. Понятие степеней свободы, пришедшее из теоретической механики, играет важную роль не только в математических дисциплинах (нам оно встретится в математической статистике), но и в классических психологических исследованиях (например, в трудах Н.А. Бернштейна).

Если попытаться представить множество решений нашей системы, то хотя бы по аналогии с линейной функцией на плоскости мы можем предположить, что множество решений представляет собой прямую линию в пространстве. Для того чтобы увидеть направления дальнейших

обобщений наших наблюдений (пока мы ограничиваемся наблюдениями, избегая доказательств), представим себе такой результат работы метода Гаусса над системой трех уравнений с тремя неизвестными

$$\left\{ x - y - z = 0. \right.$$

Преобразуем ее, перенеся часть неизвестных в правую часть

$$\left\{ x = y + z. \right.$$

В данном случае естественно будет говорить о двух степенях свободы при выборе решения: мы можем выбрать произвольно y и z, а x однозначно определится из единственного уравнения $x = y + z$. Множество решений представляет собой плоскость, проходящую через начало координат и прямые $x = y$ и $x = z$ на координатных плоскостях Oxy и Oxz соответственно.

Упражнение 1.3. Можно ли быть уверенным, что преобразования уравнений, которые используются в методе Гаусса, не приводят к потере корней и не добавляют лишних?

Упражнение 1.4. Пусть у нас есть система трех уравнений с тремя неизвестными. Можно ли быть уверенным, что при разных вариантах преобразований по методу Гаусса не получатся кардинально различные ступенчатые виды: например, при одном исходном порядке уравнений ступенчатый вид, состоящий из двух уравнений, а при другом — состоящий из трех уравнений?

1.3. Определители

В предыдущих параграфах мы фактически работали не столько с уравнениями, сколько с коэффициентами уравнений. Для сокращения записи будем использовать специальный язык таблиц чисел. Эти таблицы называются матрицами. Для системы однородных уравнений

$$\begin{cases} x + 2y + z = 0 \\ 5x + 2y + 2z = 0 \\ 3x + y + z = 0 \end{cases}$$

матрица коэффициентов будет выглядеть так:

$$\begin{pmatrix} 1 & 2 & 1 \\ 5 & 2 & 2 \\ 3 & 1 & 1 \end{pmatrix}.$$

Подробнее о матрицах речь будет идти ниже, а сейчас нам понадобится одна важная характеристика матрицы, которая называется определителем. У определителей, как и у матриц, очень много интересных употреблений, с некоторыми из которых мы познакомимся далее. Сейчас мы используем определитель, чтобы различать два класса систем линейных однородных уравнений. Сначала простейший пример:

$$\begin{cases} ax + by = 0 \\ cx + dy = 0. \end{cases}$$

Нас интересует вопрос, когда эта система будет иметь ненулевые решения? После применения метода Гаусса мы можем иметь один из двух исходов: либо получится треугольная система

$$\begin{cases} a'x + b'y = 0 \\ \qquad\; d'y = 0, \end{cases}$$

в которой a' и d' не равны 0 и которая имеет только нулевое решение (из второго уравнения $y = 0$, после чего из первого $x = 0$), либо останется только одно уравнение. Именно в последнем случае система будет иметь ненулевые решения — достаточно перенести y в правую часть и считать его свободным параметром. Очевидно, что эта ситуация возможна лишь в случае, когда исходные уравнения системы пропорциональны, т.е.

$$\frac{a}{c} = \frac{b}{d}.$$

Мы можем записать это равенство по-другому:

$$ad - bc = 0.$$

Выражение $ad-bc$ и есть интересующая нас характеристика — если $ad-bc = 0$, то наша система имеет ненулевые решения, если $ad-bc \neq 0$, то единственное решение системы $x = y = 0$.

В предыдущем рассуждении было сделано несколько ошибок. Мы легкомысленно упустили случай, когда c, или d, или оба вместе равны 0.

Упражнение 1.5. Проверить, что во всех возможных случаях соотношение $ad - bc = 0$ характеризует системы, имеющие ненулевые решения.

В разделе, посвященном факторному анализу, мы будем решать задачи о ненулевых решениях систем с очень большим количеством переменных и уравнений. Как можно было бы сформулировать единый критерий для систем любой размерности? Оказывается, таким критерием является равенство нулю определителя матрицы. У каждой квадратной матрицы можно вычислить определитель, и у матрицы размера 2×2

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

он как раз и равен $ad-bc$. Определитель матрицы A обозначается $\det A$. Определение для матриц произвольного размера мы дадим позже, а для матрицы размера 3×3 оно таково: определитель матрицы

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

мы будем обозначать также

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}$$

и вычислять по формуле $aei + dhc + bfg - gec - dbi - ahf$. Как видим, матрица заключается в круглые скобки, а ее определитель — в прямые. Заметим, что хотя обозначение для определителя похоже на обозначение матрицы, но определитель это всего лишь число, точнее, числовая характеристика матрицы.

Выражение $aei + dhc + bfg - gec - dbi - ahf$ выглядит несколько странно. Почему именно такие сочетания знаков, а не другие, вошли в определитель? Можно заметить, что во всех шести слагаемых в $aei + dhc + bfg - gec - dbi - ahf$ три сомножителя в каждом из $aei, dhc, ..., ahf$ лежат в разных строках и разных столбцах матрицы. Комбинаторное рассуждение (см. глава 4 части 3) показывает, что троек, обладающих таким свойством, ровно 6. Обратим внимание, что правило для вычисления определителя 2×2 также включает только двойки элементов матрицы, лежащих в разных строках и столбцах, — их всего две.

Что касается знаков слагаемых в сумме, то введем мнемоническое правило: со знаком "плюс" берутся такие тройки

$$\begin{array}{ccc} a & . & . \\ . & e & . \\ . & . & i \end{array} \qquad \begin{array}{ccc} . & . & c \\ d & . & . \\ . & h & . \end{array} \qquad \begin{array}{ccc} . & b & . \\ . & . & f \\ g & . & . \end{array}$$

а со знаком “минус” такие

$$\begin{array}{ccc} . & . & c \\ . & e & . \\ g & . & . \end{array} \qquad \begin{array}{ccc} a & . & . \\ . & . & f \\ . & h & . \end{array} \qquad \begin{array}{ccc} . & b & . \\ d & . & . \\ . & . & i \end{array}$$

Если буквы в каждом квадрате соединить линиями, то в первых трех будут отрезки, параллельные главной диагонали, соединяющей левый верхний угол с правым нижним, а во второй тройке, напротив, будут линии, параллельные побочной диагонали, соединяющей левый нижний угол с правым верхним. Как и в определителе 2×2, главная диагональ указывает на знак “плюс”, а побочная — на знак “минус”.

Итак, математики ручаются, что если определитель матрицы однородной системы линейных уравнений равен 0, то система имеет ненулевые решения (что очень важно психологам, которые проводят факторный анализ своих данных). Почему это так? Разобраться в этом не слишком трудно. Посмотрим, что происходит с определителем матрицы, когда мы преобразуем систему методом Гаусса. Рассмотрим четыре вида операций:

(1) перестановка строк;

(2) перестановка столбцов;

(3) умножение строки на число отличное от нуля;

(4) прибавление к строке с номером i строки с номером j, умноженной на некоторое число, с записью результата на место i-й строки.

Покажем, что если определитель не равен нулю, то эти операции переведут его также в неравный нулю определитель.

При перестановке первых двух строк, получаем матрицу

$$\begin{pmatrix} d & e & f \\ a & b & c \\ g & h & i \end{pmatrix}$$

с определителем $dbi + gec + ahf - bfg - dhc - aei$, который отличается от определителя исходной матрицы только знаками слагаемых. Точно так же прямо можно проверить, что при перестановке любых других строк, как и при перестановке столбцов, определитель матрицы меняет знак на противоположный.

Упражнение 1.6. Доказать, что определитель матрицы, имеющей две одинаковые строки, равен 0.

Рассмотрим матрицу

$$\begin{pmatrix} a & b & c \\ \lambda d & \lambda e & \lambda f \\ g & h & i \end{pmatrix},$$

где λ произвольный отличный от нуля коэффициент. Как мы отмечали, в каждом из шести слагаемых лишь один сомножитель из трех принадлежит второй строке матрицы. Это значит, что

$$\begin{vmatrix} a & b & c \\ \lambda d & \lambda e & \lambda f \\ g & h & i \end{vmatrix} = \lambda * \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix},$$

и третья операция тоже не делает ненулевой определитель нулевым.

Докажем теперь, что если у определителя имеются пропорциональные строки, то он равен нулю. Только что мы доказали, что

$$\begin{vmatrix} a & b & c \\ \lambda a & \lambda b & \lambda c \\ g & h & i \end{vmatrix} = \lambda \begin{vmatrix} a & b & c \\ a & b & c \\ g & h & i \end{vmatrix},$$

а это значит (см. упражнение 1.6), что такой определитель равен нулю.

Свойство определителя иметь в каждом слагаемом по одному представителю из каждой строки матрицы позволяет доказать еще одно необходимое нам утверждение: если строка матрицы разлагается в сумму двух строк, то и ее определитель разлагается в сумму двух определителей:

$$\begin{vmatrix} a & b & c \\ d_1 + d_2 & e_1 + e_2 & f_1 + f_2 \\ g & h & i \end{vmatrix} = \begin{vmatrix} a & b & c \\ d_1 & e_1 & f_1 \\ g & h & i \end{vmatrix} + \begin{vmatrix} a & b & c \\ d_2 & e_2 & f_2 \\ g & h & i \end{vmatrix}.$$

Определитель в левой части равен

$$a(e_1+e_2)i+h(d_1+d_2)c+b(f_1+f_2)g-g(e_1+e_2)c-b(d_1+d_2)i-a(f_1+f_2)h.$$

Раскрывая скобки и группируя слагаемые, получаем

$$ae_1i+hd_1c+bf_1g-ge_1c-bd_1i-af_1h+ae_2i+hd_2c+bf_2g-ge_2c-bd_2i-af_2h,$$

где первые 6 слагаемых составляют первый определитель в правой части, а следующие 6 — второй определитель.

В наших примерах мы рассматривали вторую строку матрицы, но ясно, что аналогичные утверждения верны и для остальных строк (и для всех столбцов).

Следствие. *Если к i-й строке матрицы прибавить j-ю строку, умноженную на любое число, то определитель матрицы не изменится.*

Действительно,

$$\begin{vmatrix} a & b & c \\ d+\lambda a & e+\lambda b & f+\lambda c \\ g & h & i \end{vmatrix}$$

разложится в сумму двух, один из которых и есть определитель исходной матрицы, а другой, имеющий пропорциональные строки, равен 0.

Мы продемонстрировали, таким образом, что если определитель исходной системы не равен нулю, то и у преобразованной методом Гаусса системы он также не равен нулю. Если же определитель равен нулю, то преобразования оставляют его равным нулю.

1.4. Определитель системы линейных уравнений ступенчатого вида

Если система

$$\begin{cases} ax + by + cz = 0 \\ dx + ey + fz = 0 \\ gx + hy + iz = 0 \end{cases}$$

после преобразований методом Гаусса превратилась в треугольную

$$\begin{cases} a'x + b'y + c'z = 0 \\ \phantom{a'x + {}} e'y + f'z = 0 \\ \phantom{a'x + b'y + {}} i'z = 0, \end{cases}$$

причем все элементы на главной диагонали (a', e' и i') отличны от нуля, то новый определитель

$$\begin{vmatrix} a' & b' & c' \\ 0 & e' & f' \\ 0 & 0 & i' \end{vmatrix}$$

будет отличен от нуля, поскольку единственное его ненулевое слагаемое есть $a'e'i'$. Но это значит, что и определитель исходной матрицы был отличен от нуля (поскольку преобразования не изменяли этого его свойства).

Если же после преобразований методом Гаусса третье уравнение исчезло, т.е. в окончательном варианте система уравнений имела вид

$$\begin{cases} a'x & + & b'y & + & c'z & = & 0 \\ & & e'y & + & f'z & = & 0 \\ & & & & 0 & = & 0, \end{cases}$$

то и определитель полученной системы

$$\begin{vmatrix} a' & b' & c' \\ 0 & e' & f' \\ 0 & 0 & 0 \end{vmatrix}$$

равен 0, поскольку равны 0 все его 6 слагаемых. Но это значит, что и определитель исходной системы был равен 0, поскольку при преобразованиях свойство “быть равным нулю” невозможно приобрести.

Теорема. *Однородная система трех линейных уравнений с тремя неизвестными имеет ненулевое решение тогда и только тогда, когда ее определитель равен* 0.

В следующей главе мы распространим наши результаты на более высокие размерности и познакомимся со строгими доказательствами.

1.5. Матрицы и векторы

Пока под вектором мы будем понимать набор чисел. Длина набора может быть разной, но пока в наших примерах будут рассматриваться только наборы длины два — $(x; y)$ и три — $(x; y; z)$. Произведение матрицы на вектор определяется следующим правилом:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} (x; y; z) = (ax + by + cz; dx + ey + fz; gx + hy + iz).$$

Упражнение 1.7. Записать правило умножения матрицы 2×2 на вектор длины 2. Можно ли столь же естественно определить произведение матрицы 2×2 на вектор длины 3?

Мы будем записывать правило умножения по-другому

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{pmatrix}.$$

В подобном произведении вектор мы представляем как столбец чисел. Такое расположение позволит нам в дальнейшем распространить эту операцию с векторов на матрицы различных размеров. Итак, квадратная матрица умножается на вектор-столбец, в результате чего получается вектор-столбец той же длины (лучше сказать, высоты).

Упражнение 1.8. Придумать правило умножения для неквадратных матриц и векторов, имеющих соответствующий размер. Можно ли умножить вектор-строку на вектор-столбец? Какому соотношению должны подчиняться их размеры?

Мы можем сказать, что матрица преобразует векторы или матрица задает преобразование векторов. При этом, чтобы найти вектор-образ v', в который перейдет данный вектор v при преобразовании, заданном матрицей A, надо матрицу умножить на вектор $v' = Av$.

Упражнение 1.9. Запишите на языке матриц и векторов следующие утверждения и проверьте, все ли они истинны.

1) Матрица, состоящая из одних нулей, даст нулевое произведение с любым вектором.

2) Умножение матрицы, у которой ненулевые элементы содержатся только в первой строке, на любой вектор даст вектор, у которого отличен от нуля только первый элемент.

3) Матрица, у которой единица стоит в верхнем левом углу, а остальные элементы — нули, при умножении на любой вектор сохранит его первую компоненту, а остальные превратит в нули.

4) Матрица, у которой на главной диагонали стоят единицы, а все остальные элементы — нули, оставляет всякий вектор без изменений.

5) Матрица, у которой на побочной диагонали стоят единицы, а все остальные элементы — нули, меняет порядок компонент вектора на противоположный.

6) Матрица 2×2, у которой строки пропорциональны, преобразует любой вектор в пропорциональный некоторому определенному вектору.

1.6. Собственные векторы

У некоторых матриц, задающих преобразование векторов, есть замечательное свойство: некоторые векторы они просто растягивают или сжимают. Такие векторы называются собственными векторами данной матрицы. В факторном анализе они и становятся факторами и интерпретируются как независимые переменные в данном процессе. Собственным вектором данной матрицы A называется вектор v, для которого $Av = \lambda v$. Число λ называется собственным значением матрицы A.

Сколько разных собственных векторов может быть у матрицы? Минимум — ни одного. Максимум — столько, каков размер матрицы. Именно столько их у симметричных матриц, с которыми имеет дело факторный анализ.

Как искать собственные векторы данной матрицы? Рассмотрим матрицу

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}.$$

Мы пока не знаем, есть ли у нее собственные векторы, тем более не знаем, каковы собственные значения, соответствующие этим векторам. Предположим, что одно такое значение есть, обозначим его буквой λ. Как найти вектор, ему соответствующий? Умножим матрицу на вектор $\begin{pmatrix} x \\ y \end{pmatrix}$ и запишем условие, говорящее о том, что этот вектор собственный с данным собственным значением.

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix}.$$

Выражение слева можно раскрыть

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + 4y \\ 2x + 3y \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix}.$$

Правое равенство означает, что $x + 4y = \lambda x$ и $2x + 3y = \lambda y$ одновременно. Таким образом, если λ собственное значение нашей матрицы, соответствующее какому-то вектору $\begin{pmatrix} x \\ y \end{pmatrix}$, то для всех этих трех неизвестных будет выполнена система равенств

$$\begin{cases} x + 4y = \lambda x \\ 2x + 3y = \lambda y \end{cases}$$

или

$$\begin{cases} x - \lambda x + 4y = 0 \\ 2x + 3y - \lambda y = 0, \end{cases}$$

причем нас интересуют ненулевые пары (x, y) ($x = y = 0$ является решением при любом λ, но нам такие нулевые векторы неинтересны).

Система содержит два уравнения с тремя неизвестными, да к тому же не является линейной, поскольку в левой части стоят произведения неизвестных. Однако мы не случайно обозначили два неизвестных латинскими буквами, а третье — греческой. Предположим, что при $\lambda = 117$ мы нашли какие-то x и y, удовлетворяющие нашей системе. Тогда

$$\begin{cases} (1 - 117)x + 4y = 0 \\ 2x + (3 - 117)y = 0, \end{cases} \tag{1.4}$$

причем, повторим, решение найдено ненулевое. Но про однородные системы линейных уравнений нам кое-что известно. Если нам удалось найти ненулевое решение, то только благодаря тому, что определитель системы равен 0, т.е.

$$\begin{vmatrix} (1 - 117) & 4 \\ 2 & (3 - 117) \end{vmatrix} = \begin{vmatrix} -116 & 4 \\ 2 & -114 \end{vmatrix} = 0.$$

117 на эту роль явно не подходит. А что подходит? Вернем на место 117 греческую букву. Подойти может только число λ, для которого

$$\begin{vmatrix} (1 - \lambda) & 4 \\ 2 & (3 - \lambda) \end{vmatrix} = 0,$$

или, раскрывая определитель, $(1 - \lambda)(3 - \lambda) - 2 * 4 = 0$. Попробуем решить полученное квадратное уравнение

$$\lambda^2 - 4\lambda - 5 = 0.$$

Равенство выполнено при двух значениях: $\lambda_1 = -1$ и $\lambda_2 = 5$. Подставим в систему (1.4) вместо 117 число 5.

$$\begin{cases} -4x + 4y = 0 \\ 2x + (-2)y = 0. \end{cases}$$

Нет ничего удивительного, что уравнения в системе пропорциональны, ведь мы искали такое значение λ, при котором определитель системы равен 0. В таком случае ненулевое решение нам гарантировано,

это, например, пара $(1;1)$ (а также пары $(2;2), (-1;-1), (-2,5;-2,5)$ и бесконечно много других, ей пропорциональных). Проверим, будет ли вектор $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ собственным. Действительно

$$\begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix},$$

то есть вектор удлинился в $\lambda = 5$ раз.

Упражнение 1.10. Проверить, что для $\lambda = -1$ также найдется собственный вектор.

Следующее упражнение требует синтезировать все знания, полученные при чтении этой главы.

Упражнение 1.11. Найти собственные векторы матрицы

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}.$$

Глава 2

Линейные уравнения (общий случай)

Мы познакомились с идеями, лежащими в основании фрагмента линейной алгебры. Некоторые из подобных вещей действительно проще понимаются на примерах, что мы и использовали в предыдущей главе, — в первую очередь это относится к идее метода Гаусса. Другие, напротив, не могут быть точно поняты без усвоения специального языка — к таким понятиям относятся, например, определители матриц 4×4 и выше. Знакомясь с данной главой, читатель освоится в этом новом языке и через некоторое время почувствует его эффективность.

Кроме того, здесь мы будем излагать материал систематически и с обычным для математических текстов структурированием, вводя определения, леммы и теоремы. Можно обратить внимание на то, что утверждения доказываются не в той минимальной формулировке, которая понадобится на следующей странице, а с той максимальной широтой, которая позволит на здесь определенное или доказанное опираться когда-нибудь в будущем — при решении неведомых еще задач. Таким образом, то, что может произвести впечатление излишества, на самом деле мотивировано стремлением при небольших текущих затратах обеспечить успех в дальнем будущем.

Изложение в этой главе совершенно параллельно изложению в предыдущей и даже названия параграфов совпадают. Поэтому здесь практически отсутствуют примеры, зато их в изобилии можно найти в параллельных местах предыдущей главы.

➽ Определение 1. *Системой n линейных уравнений с m неизвестными называется совокупность равенств*

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1m}x_m & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2m}x_m & = & b_2 \\ \dots & & \dots & & \dots & & & & \dots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nm}x_m & = & b_n \end{cases},$$

в которой a_{ij} и b_i обозначают фиксированные коэффициенты, а x_i обозначают неизвестные. Все a_{ij} и b_i могут быть равны нулю.

Система

$$\begin{cases} 0x_1 & + & 0x_2 & + & 0x_3 & = & 0 \\ 0x_1 & + & 0x_2 & + & 0x_3 & = & 0 \\ 0x_1 & + & 0x_2 & + & 0x_3 & = & 0 \end{cases}$$

вполне законна и отличается от системы

$$\begin{cases} 0x_1 & + & 0x_2 & = & 0 \\ 0x_1 & + & 0x_2 & = & 0. \end{cases}$$

Математик не будет даже спрашивать, зачем понадобилось это различение. Оно естественно, а значит, может где-то понадобиться. И действительно, отыскание собственных векторов матрицы

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

приводит к системе

$$\begin{cases} 0x_1 & + & 0x_2 & = & 0 \\ 0x_1 & + & 0x_2 & = & 0, \end{cases}$$

решениями которой оказываются произвольные пары чисел (но не тройки чисел, как было бы в случае системы 3×3), что означает, что все двумерные векторы являются собственными для данной матрицы.

2.1. Метод Гаусса

Заданную в общем виде систему уравнений можно решать методом Гаусса. Вот его алгоритм.

1. Если все коэффициенты a_{ij} системы равны нулю (см. пример выше), то система уже представлена в ступенчатом виде и работа окончена.

2. Если хотя бы один коэффициент в системе отличен от нуля, переставим строки и столбцы так, чтобы он оказался в левом верхнем углу.

Получим эквивалентную систему

$$\begin{cases} a'_{11}x'_1 & + & a'_{12}x'_2 & + & \dots & + & a'_{1m}x'_m & = & b'_1 \\ a'_{21}x'_1 & + & a'_{22}x'_2 & + & \dots & + & a'_{2m}x'_m & = & b'_2 \\ & \dots & & \dots & & \dots & & \dots & \\ a'_{n1}x'_1 & + & a'_{n2}x'_2 & + & \dots & + & a'_{nm}x'_m & = & b'_n \end{cases}.$$

Штрихи у коэффициентов и переменных поставлены потому, что мы не знаем, как именно переставлялись строки и столбцы. На месте в левом верхнем углу может оказаться коэффициент с любого другого места.

3. Поделим теперь первое уравнение на a'_{11}. Получим эквивалентную систему

$$\begin{cases} x'_1 & + & a''_{12}x'_2 & + & \dots & + & a''_{1m}x'_m & = & b''_1 \\ a'_{21}x'_1 & + & a'_{22}x'_2 & + & \dots & + & a'_{2m}x'_m & = & b'_2 \\ & \dots & & \dots & & \dots & & \dots & \\ a'_{n1}x'_1 & + & a'_{n2}x'_2 & + & \dots & + & a'_{nm}x'_m & = & b'_n \end{cases}.$$

Штрихи добавились у коэффициентов первого уравнения (поскольку они получились из предыдущих делением), но не у переменных (поскольку их порядок сохранился).

4.1. Умножим первое уравнение на a'_{21} и вычтем его из второго уравнения. Результат запишем на место второго.

...

4.$(n-1)$. Умножим первое уравнение на a'_{n1} и вычтем его из n-го уравнения. Результат запишем на место n-го.

Получим эквивалентную систему

$$\begin{cases} x'_1 & + & a''_{12}x'_2 & + & \dots & + & a''_{1m}x'_m & = & b''_1 \\ 0 & + & a''_{22}x'_2 & + & \dots & + & a''_{2m}x'_m & = & b''_2 \\ & \dots & & \dots & & \dots & & \dots & \\ 0 & + & a''_{n2}x'_2 & + & \dots & + & a''_{nm}x'_m & = & b''_n \end{cases}.$$

Если среди коэффициентов a''_{ij} в строках начиная со второй нет ни одного ненулевого, то система имеет ступенчатый вид и работа окончена. Если хотя бы один коэффициент отличен от нуля, переставим строки и столбцы так, чтобы он оказался во второй строке на втором месте. Сделаем это так, чтобы первый элемент первой строки остался на своем месте.

Получим эквивалентную систему

$$\begin{cases} x_1''' & + & a_{12}'''x_2''' & + & \dots & + & a_{1m}'''x_m''' & = & b_1''' \\ 0 & + & a_{22}'''x_2''' & + & \dots & + & a_{2m}'''x_m''' & = & b_2''' \\ \dots & & \dots & & \dots & & & & \dots \\ 0 & + & a_{n2}'''x_2''' & + & \dots & + & a_{nm}'''x_m''' & = & b_n''' \end{cases}.$$

Для единообразия мы добавили по два штриха переменным (можно было бы ограничиться одним штрихом, а первую переменную вообще оставить без изменений).

Обратим внимание на то, что последняя операция почти полностью повторяет операцию п. 1. Теперь остается перейти к п. 2, применяя его только к столбцам и строкам, начиная со вторых. Описание метода Гаусса окончено.

Довольно интересная задача — написать общий вид ступенчатой системы. Мы используем здесь тот факт, что добавление или устранение нулевого уравнения в системе, количество неизвестных которой уже определено, ничего не меняет в решениях (это значит, что системы

$$\begin{cases} 0x & + & 0y & + & 0z & = & 0 \\ 0x & + & 0y & + & 0z & = & 0 \\ 0x & + & 0y & + & 0z & = & 0 \end{cases} \text{ и } \begin{cases} 0x & + & 0y & + & 0z & = & 0 \\ 0x & + & 0y & + & 0z & = & 0 \end{cases}$$

с нашей точки зрения совершенно эквивалентны). В общем виде ступенчатая система выглядит так (штрихи мы в конце концов опускаем):

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1k}x_k & + & \dots & + & a_{1m}x_m & = & b_1 \\ 0 & + & a_{22}x_2 & + & \dots & + & a_{2k}x_k & + & \dots & + & a_{2m}x_m & = & b_2 \\ & \dots & & \dots & \dots & & & \dots & \dots & & & & \\ 0 & + & 0 & + & \dots & + & a_{kk}x_k & + & \dots & + & a_{km}x_m & = & b_k \\ 0 & + & 0 & + & \dots & + & 0 & + & \dots & + & 0 & = & b_{k+1} \\ & \dots & & \dots & \dots & & & \dots & \dots & & & & \\ 0 & + & 0 & + & \dots & + & 0 & + & \dots & + & 0 & = & b_n, \end{cases}$$

причем обязательным является условие $a_{11} \neq 0; \dots ; a_{kk} \neq 0$, где k может принимать любые значения от 0 до меньшего из чисел m и n. Внимание! Если $k = 0$, то все вообще коэффициенты a_{ij} системы равны нулю.

Понять алгоритм Гаусса в общей форме — это и значит убедиться, что заданная выше последовательность операций неизбежно приведет в конце концов к такому результату.

Свободные члены b_i в системе, приведенной к ступенчатому виду, могут оказаться какими угодно. В первую очередь существенно, есть ли ненулевые среди свободных членов $b_{k+1},\ldots,b_n$ — тех, что стоят в строках, левая часть которых нулевая. Если таковые имеются, то система несовместна, как бы хорошо ни выглядели остальные строки. Если же все эти свободные члены равны 0, то система имеет вид

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1k}x_k & + & \ldots & + & a_{1m}x_m & = & b_1 \\ 0 & + & a_{22}x_2 & + & \ldots & + & a_{2k}x_k & + & \ldots & + & a_{2m}x_m & = & b_2 \\ & \ldots & & \ldots & & \ldots & & \ldots & & \ldots & & & \\ 0 & + & 0 & + & \ldots & + & a_{kk}x_k & + & \ldots & + & a_{km}x_m & = & b_k, \end{cases}$$

который мы будем называть правильным ступенчатым видом в отличие от неправильного, соответствующего несовместным системам. Выделим еще треугольный вид

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1m}x_m & = & b_1 \\ 0 & + & a_{22}x_2 & + & \ldots & + & a_{2m}x_m & = & b_2 \\ & \ldots & & \ldots & & \ldots & & \ldots & \\ 0 & + & 0 & + & \ldots & + & a_{mm}x_m & = & b_m. \end{cases}$$

Упражнение 2.1. Являются ли системы

$$\begin{cases} 0x + y + 0z = 0 \\ 0x + 0y + 0z = 0 \end{cases} \text{и} \begin{cases} x + 0y + 0z = 0 \\ 0x + 0y + 0z = 0 \end{cases}$$

ступенчатыми?

Замечание. Мы пользовались одним непроясненным понятием и несколькими недоказанными утверждениями. Эквивалентными мы называли системы, множества решений которых совпадают. Совсем легко доказать, что к эквивалентной системе приводит перестановка строк и умножение строки на отличное от нуля число. При перестановке столбцов необходимо понимать, что новая переменная x_i''' — это одна из исходных x_j, поставленная на новое место. Говоря "эквивалентная система", надо иметь в виду, что совпадают значения x_i''' и x_j. Что касается вычитания строки, умноженной на число, из другой строки, то достаточно показать, что эта операция не приводит к потере решений (что довольно просто), а затем использовать тот факт, что прибавление той же самой строки, умноженной на то же самое число,

возвращает исходное состояние и тоже не приводит к потере решений. Строгое проведение доказательств оставляем продвинутому читателю в качестве упражнения.

Теорема 2.1. *Система треугольного вида имеет единственное решение.*

Доказательство. Пусть дана треугольная система

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1m}x_m & = & b_1 \\ 0 & + & a_{22}x_2 & + & \dots & + & a_{2m}x_m & = & b_2 \\ & \dots & & \dots & & \dots & & \dots & \\ 0 & + & 0 & + & \dots & + & a_{mm}x_m & = & b_m, \end{cases}$$

Из последнего уравнения находим $x_m = b_m / a_{mm}$ Подставляя это значение в остальные уравнения и перенося в правую часть члены, не содержащие неизвестных, получаем треугольную систему меньшего размера относительно $x_1, \dots, x_{m-1}$

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + & a_{1(m-1)}x_{m-1} & = & b_1 - a_{1m}b_m / a_{mm} \\ 0 \quad + a_{22}x_2 + \dots + & a_{2(m-1)}x_{m-1} & = & b_2 - a_{2m}b_m / a_{mm} \\ \dots \qquad \dots \qquad \dots & & \dots & \\ 0 \quad + \quad 0 \quad + \dots + & a_{(m-1)(m-1)}x_{m-1} & = & b_m - a_{(m-1)m}b_m / a_{mm}. \end{cases}$$

Повторяя данную процедуру еще необходимое число раз, получим значения всех переменных. При этом каждая переменная на соответствующем шаге получает свое значение однозначно.

Теорема доказана.

2.2. Однородные системы линейных уравнений

Рассмотрим однородную систему линейных уравнений

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1k}x_k & + & \dots & + & a_{1m}x_m & = & 0 \\ 0 & + & a_{22}x_2 & + & \dots & + & a_{2k}x_k & + & \dots & + & a_{2m}x_m & = & 0 \\ & \dots & & \dots & & \dots & & \dots & & \dots & & & \\ 0 & + & 0 & + & \dots & + & a_{kk}x_k & + & \dots & + & a_{km}x_m & = & 0. \end{cases}$$

Если $m > k$, то систему можно переписать в таком виде

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1k}x_k & = & -a_{1(k+1)}x_{k+1} & \dots - & a_{1m}x_m \\ 0 & + & a_{22}x_2 & + & \dots & + & a_{2k}x_k & = & -a_{2(k+1)}x_{k+1} & \dots - & a_{2m}x_m \\ & \dots & & \dots & & \dots & & \dots & & \dots & \\ 0 & + & 0 & + & \dots & + & a_{kk}x_k & = & -a_{m(k+1)}x_{k+1} & \dots - & a_{km}x_m. \end{cases}$$

Положим в таком случае $x_{k+1} = \ldots = x_m = 1$. По теореме 1 при заданных таким образом правых частях уравнений остальные переменные однозначно получают числовые значения, поскольку после подстановки система становится треугольной. Каковы бы ни были эти найденные значения, мы доказали, что система имеет ненулевое решение, поскольку $x_{k+1} = \ldots = x_m = 1$.

Сформулируем доказанное утверждение в виде теоремы.

Теорема 2.2. *Система правильного ступенчатого вида, у которой число неизвестных больше числа уравнений, всегда имеет ненулевое решение.*

Эта теорема уже доказана.

2.3. Определители

В предыдущей главе мы научились считать определители 2×2 и 3×3, причем последнее оказалось существенно труднее. Трудности еще более растут при переходе к высшим порядкам. Для того чтобы вычислить определитель размера 4×4, надо сложить 24 произведения четырех сомножителей, а правило выбора знаков становится достаточно сложным. К счастью, расчет определителей можно поручить компьютеру. Нам же как теоретикам понадобятся лишь некоторые их свойства, вытекающие прямо из определения.

Определение будет несколько необычным: оно будет сводить расчет определителя $n \times n$ к расчету n определителей $(n-1) \times (n-1)$. Таким образом, умея считать определитель 3×3, вы по данному правилу можете вычислить определитель 4×4 и т.д.

➽ **Определение 2**. *Алгебраическим дополнением элемента a_{ij} квадратной матрицы A, равной*

$$\begin{pmatrix} a_{11} & \cdots & a_{1(j-1)} & a_{1j} & a_{1(j+1)} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{(i-1)1} & \cdots & a_{(i-1)(j-1)} & a_{(i-1)j} & a_{(i-1)(j+1)} & \cdots & a_{(i-1)n} \\ a_{i1} & \cdots & a_{i(j-1)} & a_{ij} & a_{i(j+1)} & \cdots & a_{in} \\ a_{(i+1)1} & \cdots & a_{(i+1)(j-1)} & a_{(i+1)j} & a_{(i+1)(j+1)} & \cdots & a_{(i+1)n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{n(j-1)} & a_{nj} & a_{n(j+1)} & \cdots & a_{nn} \end{pmatrix},$$

называется определитель

$$\begin{vmatrix} a_{11} & \dots & a_{1(j-1)} & a_{1(j+1)} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{(i-1)1} & \dots & a_{(i-1)(j-1)} & a_{(i-1)(j+1)} & \dots & a_{(i-1)n} \\ a_{(i+1)1} & \dots & a_{(i+1)(j-1)} & a_{(i+1)(j+1)} & \dots & a_{(i+1)n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n(j-1)} & a_{n(j+1)} & \dots & a_{nn} \end{vmatrix},$$

умноженный на $(-1)^{(i+j)}$.

Таким образом, для того чтобы вычислить алгебраическое дополнение элемента a_{ij}, надо сначала вычеркнуть в матрице столбец и строку, содержащие данный элемент, затем вычислить определитель полученной матрицы и, наконец, сменить его знак на противоположный, если сумма номеров столбца и строки данного элемента нечетна, в противном случае оставить знак без изменения.

Алгебраическое дополнение элемента a_{ij} в матрице A обозначим $\overline{\overline{A}}_{ij}$ (две черточки над обозначающей матрицу буквой вместе с индексами намекают на то, что из матрицы вычеркнуты соответствующие строка и столбец).

Приведем примеры: алгебраическим дополнением элемента b_{22} матрицы

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

будет

$$\overline{\overline{B}}_{22} = \begin{vmatrix} b_{11} & b_{13} \\ b_{31} & b_{33} \end{vmatrix},$$

т.е. $b_{11}b_{33}-b_{31}b_{13}$, а алгебраическим дополнением элемента c_{41} матрицы

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{pmatrix}$$

будет

$$\overline{\overline{C}}_{41} = -\begin{vmatrix} c_{12} & c_{13} & c_{14} \\ c_{22} & c_{23} & c_{24} \\ c_{32} & c_{33} & c_{34} \end{vmatrix}.$$

Знак минус взят, поскольку номера столбца и строки в сумме дают 5.

Теперь мы можем дать следующее

Определение 3. *Определитель матрицы A (обозначается $\det A$) вычисляется по следующему правилу:* $\det A = a_{11}\overline{\overline{A}}_{11} + a_{21}\overline{\overline{A}}_{21} + \ldots + a_{n1}\overline{\overline{A}}_{n1}$.

Пример:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21}\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31}\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}.$$

Раскрыв определители второго порядка, получим те же самые 6 слагаемых, как и в определении предыдущей главы:

$$a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13}.$$

Если начать с определителя n-го порядка и последовательно сводить его к определителям меньшего порядка, то через n шагов получится $n!$ слагаемых, каждое из которых содержит n сомножителей, причем каждое такое произведение содержит лишь по одному представителю от каждой строки и каждого столбца. Выбор знака для произведения диктуется следующим правилом, которое мы изобразим наглядно. Пусть имеется следующее произведение.

$$\begin{vmatrix} c_{11} & \cdot & \cdot & \cdot \\ \cdot & \cdot & c_{23} & \cdot \\ \cdot & \cdot & \cdot & c_{34} \\ \cdot & c_{42} & \cdot & \cdot \end{vmatrix}$$

Для определения того, с каким знаком надо включить его в сумму, будем считать для каждого его сомножителя, сколько сомножителей находится одновременно выше и правей него. Для c_{11} таких нет, для c_{23} тоже нет, для c_{34} тоже нет, но для c_{42} их сразу два. В сумме $0+0+0+2$ дают четное число 2, что говорит о том, что в сумме для подсчета определителя это произведение надо взять со знаком "плюс". Если получается нечетная сумма, то перед этим произведением ставим знак "минус". Это правило легко приложимо к определителям любых порядков.

Назовем этот способ вычисления определителя "Правилом знаков".

Теорема 2.3. *Вычисление определителя способом, данным в Определении 3, и способом, данным Правилом знаков, приводит к одному и тому же результату.*

Громоздкое доказательство теоремы мы опускаем. В дальнейшем мы будем апеллировать в рассуждениях к одному или другому правилу вычисления из соображений удобства.

Свойства определителей

1. *Если отразить матрицу относительно главной диагонали (такая операция называется транспонированием), ее определитель не изменится.*

Доказательство. Сравним состоящие из одних и тех же сомножителей соответствующие произведения в исходной и транспонированной матрицах:

$$\begin{vmatrix} a & . & . & . \\ . & . & b & . \\ . & . & . & c \\ . & d & . & . \end{vmatrix} \text{ и } \begin{vmatrix} a & . & . & . \\ . & . & . & d \\ . & b & . & . \\ . & . & c & . \end{vmatrix}$$

Заметим, что если в паре элементов один находится “выше и правее”, то другой — “ниже и левее”, поэтому правило знаков можно было бы сформулировать и так: “будем считать для каждого сомножителя, сколько сомножителей находится одновременно ниже и левее него”. Но при транспонировании элемент, находящийся “выше и правее” оказывается “ниже и левее”, а это значит, что если для исходной матрицы определять знак по первому варианту правила знаков (“выше и правее”), а для транспонированной — по второму (“ниже и левее”), то для каждого элемента данного произведения результаты совпадут (в приведенном выше примере в исходной матрице два элемента лежат выше и правее элемента d, в транспонированной, которая ей симметрична, эти же два элемента лежат ниже и левее d). **Утверждение доказано**.

2а. *Если в матрице поменять местами две соседние строки, то абсолютная величина ее определителя не изменится, а знак поменяется на противоположный.*

Доказательство. Рассмотрим некоторое произведение сначала в исходном определителе, а затем в полученном из него перестановкой 2-й и 3-й строк.

$$\begin{vmatrix} a & . & . & . \\ . & . & b & . \\ . & . & . & c \\ . & d & . & . \end{vmatrix} \text{ и } \begin{vmatrix} a & . & . & . \\ . & . & . & c \\ . & . & b & . \\ . & d & . & . \end{vmatrix}$$

Если в исходном определителе пара сомножителей, принадлежащих этим строкам, не состояла в отношении “один выше-правее, другой ниже-левее” (пара элементов b и c в левом определителе), то при перестановке строк они это отношение приобретут (см. правый определитель), поэтому сумма, вычисляемая для принятия решения по правилу знаков, окажется на единицу больше.

Наоборот, если исходно пара состояла в данном отношении, то после перестановки строк она его утратит, и сумма уменьшится на единицу.

В любом случае знак перед данным произведением поменяется на противоположный, и это будет верно для каждого произведения, входящего в определитель, — поскольку каждое произведение имеет ровно по одному элементу в каждой из этих строк.

Подобные рассуждения можно провести для любой пары строк, поэтому **свойство 2а доказано**.

2б. *Если в матрице поменять местами любые две строки, то абсолютная величина ее определителя не изменится, а знак поменяется на противоположный.*

Доказательство этого факта, собственно говоря, не принадлежит линейной алгебре. Пусть номера переставленных строк i и $i+k$. Пусть к тому же нам временно разрешается менять местами только соседние строки. Посчитаем, сколько потребуется таких мелких операций, чтобы достичь конечного результата — поменять местами строки i и $i+k$. Сначала будем менять местами строку i с нижними соседними k раз, пока она не окажется на месте $i+k$. При последней перестановке одновременно вторая из наших строк переместится с $(i+k)$-го места на $(i+k-1)$-е. Теперь $k-1$ перестановка этой строки с верхними ее ближайшими соседями поставит ее на мето i. Все остальные строки вернутся на свои исходные места, а мы за $k+k-1$ перестановок соседних строк произвели требовавшуюся операцию. Число $k+k-1=2k-1$ всегда нечетное, при каждой из $2k-1$ операций знак менялся на противоположный, в итоге он окажется противоположным исходному. **Свойство 2б доказано**.

2в. *Если в матрице поменять местами любые два столбца, знак ее определителя поменяется на противоположный.*

Доказательство. Транспонируем матрицу, поменяем местами соответствующие строки и затем опять транспонируем полученную матрицу.

В результате столбцы поменяются местами. По свойству 1 оба транспонирования не меняют знак, по свойству 2б при перестановке строк знак меняется на противоположный. **Свойство 2в доказано.**

2. *При перестановке любых двух строк или столбцов матрицы абсолютная величина ее определителя сохраняется, а знак меняется на противоположный.*

Это свойство представляет собой резюме свойств 2б и 2в.

3. *Умножение строки или столбца матрицы на некоторое число λ приводит к умножению ее определителя на то же самое число λ. Если матрица имеет нулевую строку (столбец), то ее определитель равен нулю.*

Доказательство. В каждом из $n!$ произведений, входящих в определитель, имеется ровно один сомножитель из измененной строки, т.е. каждое произведение умножается на λ, а значит, и вся сумма, составляющая определитель, умножается на λ. Если $\lambda = 0$, то и определитель равен нулю.

4. *Если две строки или столбца матрицы пропорциональны, то ее определитель равен нулю.*

Доказательство. Если две строки (столбца) равны, то определитель равен нулю, поскольку перестановка этих строк (столбцов) должна поменять знак определителя, а при этой перестановке матрица не меняется, следовательно, ее определитель равен нулю. Для пропорциональных строк (столбцов) надо сначала применить свойство 3, затем то же самое рассуждение.

5а. *Если первая строка матрицы раскладывается в сумму двух строк, то и ее определитель раскладывается в сумму соответствующих определителей.*

Доказательство. Пусть первая строка данного определителя представляет собой сумму двух строк:

$$\begin{vmatrix} a'_{11}+a''_{11} & a'_{12}+a''_{12} & \dots & a'_{1n}+a''_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Для вычисления определителя воспользуемся определением 3.

$$\begin{vmatrix} a'_{11}+a''_{11} & a'_{12}+a''_{12} & \dots & a'_{1n}+a''_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} =$$

$$= (a'_{11}+a''_{11})\overline{\overline{A}}_{11} + (a'_{12}+a''_{12})\overline{\overline{A}}_{12} + \ldots + (a'_{1n}+a''_{1n})\overline{\overline{A}}_{1n}.$$

Раскрыв скобки и произведя перегруппировку, получаем

$$(a'_{11}+a''_{11})\overline{\overline{A}}_{11} + (a'_{12}+a''_{12})\overline{\overline{A}}_{12} + \ldots + (a'_{1n}+a''_{1n})\overline{\overline{A}}_{1n} =$$

$$= (a'_{11}\overline{\overline{A}}_{11} + a'_{12}\overline{\overline{A}}_{12} + \ldots + a'_{1n}\overline{\overline{A}}_{1n}) + (a''_{11}\overline{\overline{A}}_{11} + a''_{12}\overline{\overline{A}}_{12} + \ldots + a''_{1n}\overline{\overline{A}}_{1n}).$$

Выражения в скобках по определению 3 представляют собой искомые определители

$$\begin{vmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} + \begin{vmatrix} a''_{11} & a''_{12} & \dots & a''_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Свойство 5а доказано.

5. *Если некоторая строка (столбец) матрицы раскладывается в сумму двух строк (столбцов), то и ее определитель раскладывается в сумму соответствующих определителей.*

Доказательство для строк. Пусть данная строка имеет номер i. Поменяем ее местами с первой строкой (ее определитель умножится на -1), затем разложим в сумму двух определителей и затем опять поменяем в обоих полученных первую и i-ю строки (определители умножатся еще раз на -1). Но $(-1) * (-1) = 1$, тем самым **утверждение доказано**.

Доказательство для столбцов. Транспонируем матрицу, применим предыдущее рассуждение и опять транспонируем обе матрицы. **Утверждение доказано.**

6а. *Прибавление к первой строке матрицы какой-то другой ее j-й строки ($j \neq 1$), умноженной на число λ, не меняет ее определителя.*

Доказательство. Нам надо посчитать определитель, в котором к первой строке прибавлена j-я, умноженная на число λ:

$$\begin{vmatrix} a_{11}+\lambda a_{j1} & a_{12}+\lambda a_{j2} & \dots & a_{1n}+\lambda a_{jn} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

По свойству 5 он раскладывается в сумму двух определителей, в которой первый есть исходный определитель, а второй имеет пропорциональные строки и по свойству 4 равен нулю:

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} + \begin{vmatrix} \lambda a_{j1} & \lambda a_{j2} & \dots & \lambda a_{jn} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

6. *Прибавление к i-й строке матрицы ее j-й строки ($j \neq i$), умноженной на произвольное число λ, не меняет ее определителя.*

Доказательство. Как и в доказательстве свойства 5, надо сначала поменять местами i-ю строку с первой, доказать равенство и вернуть строки на место.

7. *Определитель матрицы, в которой ниже главной диагонали стоят только нули, а на главной диагонали отличные от нуля числа, равен произведению диагональных элементов.*

Доказательство. Пусть дан определитель

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1(n-1)} & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2(n-1)} & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3(n-1)} & a_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{(n-1)(n-1)} & a_{(n-1)n} \\ 0 & 0 & 0 & \dots & 0 & a_{nn} \end{vmatrix}.$$

Единственное не содержащее нулей произведение, включающее по одному представителю из каждой строки и каждого столбца, — это произведение диагональных элементов. Действительно, из первого столбца можно взять только a_{11}, но это значит, что из второго столбца — только a_{22}, поскольку первая строка уже представлена первым элементом, и т.д. до a_{nn}. Таким образом, искомый определитель равен $a_{11}a_{22}a_{33}\ldots a_{(n-1)(n-1)}a_{nn}$.

2.4. Определитель системы линейных уравнений ступенчатого вида

Теперь мы докажем в общем виде теорему о связи определителя и наличия ненулевых решений системы линейных уравнений, у которой число неизвестных равно числу уравнений.

Теорема 2.4. *Если определитель однородной системы равен нулю, то она имеет ненулевые решения, если же определитель не равен нулю, то решение системы только нулевое.*

Доказательство, как и в предыдущем параграфе, будет использовать метод Гаусса. Покажем сначала, что элементарные преобразования строк и столбцов, которые мы будем использовать, не меняют важного для нас свойства определителя системы.

1) Деля некоторое уравнение в системе на отличное от нуля число, мы делим на это же самое число и строку определителя, а значит, и сам определитель системы (свойство 3 определителя). Если определитель системы был равен нулю, то он останется таковым, если не был равным нулю, то останется отличным от нуля.

2) Переставляя строки или столбцы, мы меняем только знак определителя.

3) Прибавляя или вычитая из одного уравнения другое, умноженное на число, мы оставляем определитель без изменения.

Таким образом, приведя систему к ступенчатому виду, мы неравный нулю определитель оставляем неравным, а равный нулю — равным.

Пусть теперь дана однородная система линейных уравнений общего вида, у которой количество уравнений равно количеству неизвестных.

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & 0 \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & 0 \\ & \ldots & & \ldots & & \ldots & & & \ldots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & 0. \end{cases}$$

В результате ее преобразований методом Гаусса мы получим ступенчатую систему

$$\begin{cases} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1k}x_k & + & \dots & + & a_{1n}x_n & = & 0 \\ 0 & + & a_{22}x_2 & + & \dots & + & a_{2k}x_k & + & \dots & + & a_{2n}x_n & = & 0 \\ & \dots & & \dots & & \dots & & \dots & & \dots & & & \\ 0 & + & 0 & + & \dots & + & a_{kk}x_k & + & \dots & + & a_{kn}x_n & = & 0 \\ 0 & + & 0 & + & \dots & + & 0 & + & \dots & + & 0 & = & 0 \\ & \dots & & \dots & & \dots & & \dots & & \dots & & & \\ 0 & + & 0 & + & \dots & + & 0 & + & \dots & + & 0 & = & 0. \end{cases}$$

Если $k < n$, т.е. последние несколько строк системы нулевые, то

1) определитель системы с самого начала равнялся нулю, поскольку он оказался равным нулю после гауссовских преобразований;

2) система имеет ненулевые решения (теорема 2).

Если $k = n$, т.е. система имеет треугольный вид, то

1) определитель системы не равен нулю, поскольку на главной диагонали стоят ненулевые элементы, а значит, и с самого начала он не был равен нулю;

2) система имеет единственное решение, а именно нулевое (теорема 1).

Теорема доказана.

2.5. Матрицы и векторы

➽ **Почти определение 4.** *n чисел, записанные в столбик и заключенные в круглые скобки, мы будем называть n-мерным вектором (в дальнейшем понятие вектора будет расширяться).*

➽ **Определение 5**. *Красивая таблица размера $m \times n$ чисел или алгебраических выражений, заключенная в круглые скобки, называется матрицей (уговоримся считать m горизонтальным размером, а n вертикальным).*

Из определений следует, что матрицу размера $1 \times n$ можно называть вектором.

Матрицы можно умножать на векторы, если горизонтальный размер матрицы равен длине вектора.

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nk} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_k \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \dots \\ \dots \\ c_n \end{pmatrix},$$

где $c_i = a_{i1}b_1 + a_{i2}b_2 + \ldots + a_{ik}b_k$.

Обратим внимание, что длина вектора-множителя равна ширине матрицы, а вектор-результат наследует высоту матрицы.

Упражнение 2.2. Запишите на языке матриц и векторов общего вида следующие утверждения и проверьте, все ли они истинны.

1) Матрица, состоящая из одних нулей, даст нулевое произведение с любым вектором.

2) Умножение матрицы, у которой ненулевые элементы содержатся только в первой строке, на любой вектор даст вектор, у которого отличным от нуля может быть только первый элемент.

3) Матрица, у которой единица стоит в верхнем левом углу, а остальные элементы — нули, при умножении на любой вектор сохранит его первую компоненту, а остальные превратит в нули.

4) Квадратная матрица, у которой на главной диагонали стоят единицы, а все остальные элементы — нули, оставляет всякий вектор без изменений.

5) Матрица, у которой на побочной диагонали стоят единицы, а все остальные элементы — нули, меняет порядок компонент вектора на противоположный.

6) Матрица $n \times n$, у которой строки пропорциональны, преобразует любой вектор в пропорциональный некоторому определенному вектору.

2.6. Собственные векторы

➽ **Определение 6.** *Два вектора*

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix} \text{ и } w = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{pmatrix}$$

равны между собой, если $v_1 = w_1$; $v_2 = w_2$; $\ldots$; $v_n = w_n$.
Вектор

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \ldots \\ v_n \end{pmatrix}$$

равен нулю, если равны нулю все его компоненты, т.е. $v_1 = v_2 = \ldots = v_n = 0$.

➽ **Определение 7**. *Собственным вектором квадратной матрицы A размера $n \times n$ называется отличный от нуля вектор v длины n, если для некоторого числа λ (которое может быть равным нулю) $Av = \lambda v$. Число λ называется собственным значением матрицы.*

Пусть даны матрица и вектор

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}, \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \ldots \\ v_n \end{pmatrix}.$$

Поскольку

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \ldots \\ v_n \end{pmatrix} = \begin{pmatrix} a_{11}v_1 + a_{12}v_2 + \ldots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \ldots + a_{2n}v_n \\ \ldots \\ a_{n1}v_1 + a_{n2}v_2 + \ldots + a_{nn}v_n \end{pmatrix}$$

и именно этот, стоящий в предыдущем равенстве справа от знака равенства вектор и должен быть равен вектору

$$\begin{pmatrix} \lambda v_1 \\ \lambda v_2 \\ \ldots \\ \lambda v_n \end{pmatrix},$$

то для v “быть собственным вектором, соответствующим собственному значению λ” означает быть решением системы уравнений

$$\begin{cases} a_{11}v_1 + a_{12}v_2 + \ldots + a_{1n}v_n = \lambda v_1 \\ a_{21}v_1 + a_{22}v_2 + \ldots + a_{2n}v_n = \lambda v_2 \\ \ldots \\ a_{n1}v_1 + a_{n2}v_2 + \ldots + a_{nn}v_n = \lambda v_n \end{cases}$$

или однородной системы

$$\begin{cases} (a_{11}-\lambda)v_1 + a_{12}v_2 + \ldots + a_{1n}v_n = 0 \\ a_{21}v_1 + (a_{22}-\lambda)v_2 + \ldots + a_{2n}v_n = 0 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ a_{n1}v_1 + a_{n2}v_2 + \ldots + (a_{nn}-\lambda)v_n = 0. \end{cases}$$

По теоремам 1 и 2 такая система может иметь ненулевое решение, только если равен нулю определитель системы. Таким образом, если мы хотим найти собственный вектор матрицы, мы можем начать с поиска ее собственного значения, которое должно удовлетворять простому уравнению, называемому характеристическим уравнением,

$$\begin{vmatrix} (a_{11}-\lambda) & a_{12} & \ldots & a_{1k} \\ a_{21} & (a_{22}-\lambda) & \ldots & a_{1k} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & (a_{nn}-\lambda) \end{vmatrix} = 0$$

(напомним, что определитель — это число или заменяющее число алгебраическое выражение). Затем с найденным собственным значением нам следует решить систему уравнений

$$\begin{cases} (a_{11}-\lambda)v_1 + a_{12}v_2 + \ldots + a_{1n}v_n = 0 \\ a_{21}v_1 + (a_{22}-\lambda)v_2 + \ldots + a_{2n}v_n = 0 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ a_{n1}v_1 + a_{n2}v_2 + \ldots + (a_{nn}-\lambda)v_n = 0, \end{cases}$$

которая даст нам соответствующий собственный вектор, причем тот факт, что определитель системы после подстановки найденного значения λ равен нулю, гарантирует наличие ненулевого собственного вектора в качестве решения системы уравнений.

Пример. Найдем собственные векторы матрицы

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 3 & 3 & 3 & 4 \\ 4 & 4 & 3 & 4 \end{pmatrix}$$

Ищем собственные значения, решая уравнение

$$\begin{vmatrix} 1-\lambda & 0 & 0 & 0 \\ 2 & 2-\lambda & 0 & 0 \\ 3 & 3 & 3-\lambda & 4 \\ 4 & 4 & 3 & 4-\lambda \end{vmatrix} = 0.$$

Считать определитель будем по определению 3, разложив его по первой строке: $\det A = a_{11}\overline{\overline{A}}_{11} + a_{12}\overline{\overline{A}}_{12} + a_{13}\overline{\overline{A}}_{13} + a_{14}\overline{\overline{A}}_{14}$.

К счастью, $a_{12} = a_{13} = a_{14} = 0$, поэтому остается только

$$\det A = (1-\lambda)\begin{vmatrix} 2-\lambda & 0 & 0 \\ 3 & 3-\lambda & 4 \\ 4 & 3 & 4-\lambda \end{vmatrix}.$$

По тому же правилу разложения по строке оставшийся определитель в этой формуле можно выразить так:

$$\begin{vmatrix} 2-\lambda & 0 & 0 \\ 3 & 3-\lambda & 4 \\ 4 & 3 & 4-\lambda \end{vmatrix} = (2-\lambda)\begin{vmatrix} 3-\lambda & 4 \\ 3 & 4-\lambda \end{vmatrix} = (2-\lambda)[(3-\lambda)(4-\lambda)-12].$$

И окончательное уравнение

$$\det A = (1-\lambda)(2-\lambda)[(3-\lambda)(4-\lambda)-12] = 0$$

после преобразования содержимого квадратных скобок принимает вид

$$(1-\lambda)(2-\lambda)[(\lambda-7)\lambda] = 0.$$

Возьмем одно из решений уравнения, $\lambda = 7$. Для того чтобы найти собственный вектор, нам предстоит решить систему

$$\begin{cases} (1-7)v_1 & & & & = 0 \\ 2v_1 + (2-7)v_2 & & & & = 0 \\ 3v_1 + 3v_2 + (3-7)v_3 + 4v_4 & & & & = 0 \\ 4v_1 + 4v_2 + 3v_3 + (4-7)v_4 & & & & = 0. \end{cases}$$

Даже не считая разности в скобках, мы можем сразу увидеть, что первые два уравнения дают $v_1 = v_2 = 0$. Подставив эти значения в последние два уравнения и произведя вычитание в скобках, получаем эквивалентную систему:

$$\begin{cases} v_1 = 0 \\ v_2 = 0 \\ -4v_3 + 4v_4 = 0 \\ 3v_3 - 3v_4 = 0, \end{cases}$$

Нет ничего удивительного в том, что третье и четвертое уравнения пропорциональны, поскольку определитель системы изначально равен

нулю (именно такое значение параметра λ мы и искали). Решением системы будет, например, вектор

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Хотя процедура гарантирует результат, проверим, что данный вектор — собственный.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 3 & 3 & 3 & 4 \\ 4 & 4 & 3 & 4 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 3*0+3*0+3*1+4*1 \\ 4*0+4*0+3*1+4*1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 7 \\ 7 \end{pmatrix}.$$

Мы убедились, что наш вектор действительно "растянулся" в 7 раз.

Упражнение 2.3.

1) Какие еще векторы соответствуют собственному значению $\lambda = 7$? Проверить умножением матрицы на вектор.

2) Найти собственные векторы, соответствующие остальным трем корням характеристического уравнения из разобранного выше примера.

Глава 3

Векторы и матрицы (аналитическая геометрия)

В школьном курсе геометрии вектор определялся как направленный отрезок на плоскости или в трехмерном евклидовом пространстве. Плоскость и трехмерное пространство, хотя и являются идеальными объектами математики, но они даны нам в некоторой непосредственной наглядной интуиции, подкрепляемой возможностью изображать исследуемые предметы на чертежах.

Наше данное в предыдущей главе "почти определение" вектора как набора чисел было чисто алгебраическим, все проведенные рассуждения не апеллировали ни к какой наглядности, кроме наглядности алгебраических тождеств. Какое определение более правильное, и что же такое вектор?

➽ **Не определение 1**. *Вектор это и направленный отрезок, и набор чисел.*

Упражнение 3.1.

1) представить себе направленный отрезок в трехмерном пространстве;

2) представить себе направленный отрезок в четырехмерном пространстве.

Первое задание упражнения 3.1 удается сделать без труда. Второе задание невыполнимо для человеческих существ, если под словами "представить себе" понимать обычное человеческое воображение.

Однако психологам приходится иметь дело с многомерными пространственными конфигурациями (например, пространство десяти основных шкал теста MMPI) и предполагать, что с их помощью описывается какая-то реальность (в чем-то аналогичная реальности населяемого нами пространства, которое описывается геометрией "направленных отрезков"). Как же представлять себе эту реальность?

Смысл синтеза, который проводит линейная алгебра, состоит в том, чтобы единым образом описывать ситуации в "пространствах" произвольного числа измерений. В настоящей главе мы произведем совмещение алгебраического и геометрического понимания вектора на плоскости, а в следующей главе займемся векторными пространствами в общем случае произвольного числа измерений.

3.1. Векторы в двумерном пространстве

Будем считать векторами на плоскости направленные отрезки, начала которых лежат в некоторой общей точке O.

Из школьного курса известно, что направленные отрезки на плоскости можно складывать и умножать на действительные числа. Для того чтобы сложить векторы $\mathbf{v}$ и $\mathbf{w}$, надо построить параллелограмм, две стороны которого образуют векторы $\mathbf{v}$ и $\mathbf{w}$, и построить направленный отрезок с началом в точке O и с концом в противоположной вершине параллелограмма (рис. 3.1).

Для того чтобы умножить вектор $\mathbf{v}$ на число λ, надо

— удлинить его в λ раз, если λ положительное число,

— удлинить в $|\lambda|$ раз и отразить относительно точки O, если λ отрицательное число.

Если $\lambda = 0$, то результатом умножения будет нулевой вектор с началом и концом в одной и той же точке O.

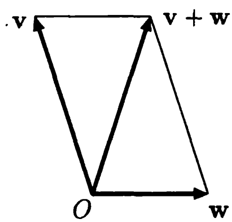


Рис. 3.1

Важным свойством векторов на плоскости является возможность разложения вектора по базису. Любые два не лежащих на одной прямой вектора могут считаться базисом на плоскости. Чтобы разложить вектор $\mathbf{s}$ по данному базису $\{\mathbf{u}, \mathbf{v}\}$, надо построить параллелограмм, у которого две противоположные вершины — точка O и конец вектора $\mathbf{s}$, а стороны параллельны векторам $\mathbf{u}$ и $\mathbf{v}$ (рис. 3.2). Если другие две вершины считать концами векторов $\mathbf{u}'$ и $\mathbf{v}'$, то $\mathbf{s} = \mathbf{u}' + \mathbf{v}'$ по правилу

параллелограмма. Но $\mathbf{u}'$ лежит на одной прямой с $\mathbf{u}$, поэтому можно найти число λ, такое, что $\mathbf{u}' = \lambda\mathbf{u}$, и точно так же можно найти число μ, такое, что $\mathbf{v}' = \mu\mathbf{v}$. Таким образом, $\mathbf{s} = \lambda\mathbf{u} + \mu\mathbf{v}$. Это и значит, что вектор $\mathbf{s}$ разложен по базису $\{\mathbf{u}, \mathbf{v}\}$.

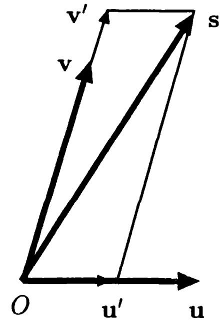


Рис. 3.2

Заметим, что мы не требуем, чтобы векторы базиса были взаимно перпендикулярны. Если читателю трудно следить за общим изложением, он может рассматривать только ортонормированные базисы, т.е. такие, в которых базисные векторы перпендикулярны и имеют единичную длину. В большинстве своем формулы этой главы верны как для этих базисов, так и в общем виде.

3.2. Линейные преобразования

Новая важная операция над векторами в линейной алгебре — это линейное преобразование. Линейное преобразование $\mathbf{A}$ ставит в соответствие каждому вектору $\mathbf{u}$ некоторый другой вектор $\mathbf{u}' = \mathbf{A}(\mathbf{u})$, причем

1) если $\mathbf{v} = \lambda\mathbf{u}$,
$\mathbf{u}' = \mathbf{A}(\mathbf{u})$ и $\mathbf{v}' = \mathbf{A}(\mathbf{v})$,
то $\mathbf{v}' = \lambda\mathbf{u}'$;

2) если $\mathbf{u} = \mathbf{v} + \mathbf{w}$,
$\mathbf{u}' = \mathbf{A}(\mathbf{u})$, $\mathbf{v}' = \mathbf{A}(\mathbf{v})$ и $\mathbf{w}' = \mathbf{A}(\mathbf{w})$,
то $\mathbf{u}' = \mathbf{v}' + \mathbf{w}'$.

Вектор $\mathbf{u}' = \mathbf{A}(\mathbf{u})$ называется образом вектора $\mathbf{u}$ при преобразовании $\mathbf{A}$. Вышеприведенные свойства можно переформулировать так:

— образ вектора, умноженного на некоторое число, равен умноженному на это же число образу данного вектора;

— образ суммы двух векторов равен сумме образов этих векторов.

Можно выразить оба свойства одной формулой:

$$\mathbf{A}(\lambda\mathbf{u} + \mu\mathbf{v}) = \lambda\mathbf{A}(\mathbf{u}) + \mu\mathbf{A}(\mathbf{v}) \tag{3.1}$$

Примеры линейных преобразований:

1) все векторы отражаются относительно точки O;

2) все векторы поворачиваются по часовой стрелке вокруг точки O на прямой угол и удлиняются в два раза;

3) все векторы отображаются в нулевой вектор;

4) все векторы отражаются относительно некоторой прямой, проходящей через точку O.

Примеры преобразований, которые не являются линейными:

1) ко всем векторам прибавляется один и тот же фиксированный вектор;

2) все векторы отображаются в один и тот же фиксированный вектор.

3.3. Связь преобразования, базиса и матрицы

Выберем на плоскости некоторый базис $\{\mathbf{e_1}, \mathbf{e_2}\}$ и рассмотрим следующее линейное преобразование $\mathbf{A}$:

$\mathbf{A}(\mathbf{e_1}) = \mathbf{e_1}$, $\mathbf{A}(\mathbf{e_2}) = \mathbf{0}$.

Этих двух равенств и линейности преобразования достаточно, чтобы узнать образ любого вектора. Любой вектор $\mathbf{v}$ можно разложить по базису: $\mathbf{v} = \lambda_1\mathbf{e_1} + \lambda_2\mathbf{e_2}$.

По характеризующей линейный оператор формуле (3.1)

$$\mathbf{A}(\mathbf{v}) = \mathbf{A}(\lambda_1\mathbf{e_1} + \lambda_2\mathbf{e_2}) = \lambda_1\mathbf{A}(\mathbf{e_1}) + \lambda_2\mathbf{A}(\mathbf{e_2}).$$

Прежде чем подставить в формулу известные значения, заметим, что она показывает, что любое линейное преобразование векторов плоскости вполне определяется образами двух своих базисных векторов (т.е. любых двух неколинеарных векторов).

Поскольку для нашего оператора $\mathbf{A}(\mathbf{e_2}) = \mathbf{0}$,

$$\mathbf{A}(v) = \lambda_1\mathbf{A}(\mathbf{e_1}) + \lambda_2\mathbf{A}(\mathbf{e_2}) = \lambda_1\mathbf{A}(\mathbf{e_1}),$$

т.е. (см. рис. 3.3) образ вектора $\mathbf{v}$ равен проекции $\mathbf{v}$ на вектор $\mathbf{e_1}$ (параллельно $\mathbf{e_2}$ в неортогональном случае).

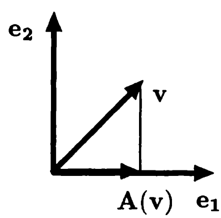


Рис. 3.3. Пример линейного преобразования

Теперь мы свяжем с преобразованием и базисом матрицу. Всякий вектор плоскости разложим по нашему базису

$$\mathbf{v} = x\mathbf{e_1} + y\mathbf{e_2}.$$

Мы будем теперь именовать вектор **v** записанной в столбик парой чисел

$$\begin{pmatrix} x \\ y \end{pmatrix}$$

(обнаружив тем самым связь с "почти определением" вектора, данным в предыдущей главе). Эти числа называют координатами вектора в данном базисе. Их принято обозначать латинскими буквами x, y, $z, \ldots$ (с индексами или без).

Несколько слов о языке. В этой и последующих главах векторы как "реально" существующие в каком-то пространстве "вещи" мы будем обозначать "жирным" шрифтом. Векторы-столбцы будут всегда связаны с обычными в математических формулах тонкими курсивными буквами.

Упражнение 3.2. Какие имена получат векторы $\mathbf{e}_1$ и $\mathbf{e}_2$?

Ответ: $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ и $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ соответственно.

В новом языке наше преобразование **A** будет отображать вектор с именем $\begin{pmatrix} x \\ y \end{pmatrix}$ в вектор с именем $\begin{pmatrix} x \\ 0 \end{pmatrix}$.

Теперь, наконец, мы свяжем с преобразованием **A** матрицу

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

и убедимся, что именно эта матрица производит данное преобразование над векторами-столбцами:

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix}.$$

Матрица преобразования не была угадана. Посмотрим, как произвольная матрица действует на базисные векторы (см. упражнение 3.2):

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} \text{ и } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}.$$

Образы базисных векторов стали столбцами матрицы. Таким образом, для того чтобы задать матрицу преобразования (в данном базисе), надо знать образы (в том же базисе) базисных векторов, которые следует записать в качестве столбцов данной матрицы.

Мы поставили в соответствие каждому направленному отрезку на плоскости пару чисел, которые являются его проекциями на базисные векторы. Линейному преобразованию направленных отрезков на плоскости мы поставили в соответствие матрицу. Мы называем эти векторы-столбцы и эту матрицу именами векторов-стрелок (направленных отрезков) и их преобразования, поскольку в другом каком-нибудь базисе те же направленные отрезки и то же преобразование будут иметь другие выражения в виде столбцов и матрицы.

Пример 1.

Для пояснения рассмотрим на той же плоскости, что и в предыдущем примере, новый базис $\{\mathbf{f}_1, \mathbf{f}_2\}$, состоящий из векторов $\mathbf{f}_1 = \mathbf{e}_2$; $\mathbf{f}_2 = \mathbf{e}_1$.

Мы имеем в виду, что все векторы-стрелки на плоскости остались на своих местах, а базис берется новый. Векторы-стрелки получают новые имена: тот, что имел, например, имя $\begin{pmatrix} 17 \\ 2 \end{pmatrix}$ получает имя $\begin{pmatrix} 2 \\ 17 \end{pmatrix}$.

При нашем линейном преобразовании этот вектор отображается в вектор, имеющий в первом базисе имя $\begin{pmatrix} 17 \\ 0 \end{pmatrix}$, а в новом имя $\begin{pmatrix} 0 \\ 17 \end{pmatrix}$.

Матрица нашего преобразования также будет в новом базисе иной. По ее столбцам следует записать координаты в новом базисе новых базисных векторов. Поскольку $\mathbf{A}(\mathbf{e}_1) = \mathbf{e}_1$, $\mathbf{A}(\mathbf{e}_2) = \mathbf{0}$ (не забудем, что буква $\mathbf{A}$ обозначает у нас преобразование векторов-стрелок на плоскости), то $\mathbf{A}(\mathbf{f}_1) = \mathbf{0}$, а $\mathbf{A}(\mathbf{f}_2) = \mathbf{f}_2$.

Это значит, что первый столбец новой матрицы должен быть нулевым, а во втором на первом месте нуль, а на втором единица. Наше преобразование получает в новом базисе новую матрицу

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

При умножении на новое имя $\begin{pmatrix} 2 \\ 17 \end{pmatrix}$ это дает

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 17 \end{pmatrix} = \begin{pmatrix} 0 \\ 17 \end{pmatrix},$$

т.е. новое имя образа.

Подведем итог. Тот самый вектор $\mathbf{v}$, который имел в двух базисах разные имена $\begin{pmatrix} 17 \\ 2 \end{pmatrix}$ и $\begin{pmatrix} 2 \\ 17 \end{pmatrix}$, преобразуется в вектор $\mathbf{v}'$, что можно

записать равенством для векторов-стрелок

$\mathbf{A}(\mathbf{v}) = \mathbf{v}'$.

В первом базисе это равенство приобретает вид

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 17 \\ 2 \end{pmatrix} = \begin{pmatrix} 17 \\ 0 \end{pmatrix},$$

а во втором

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 17 \end{pmatrix} = \begin{pmatrix} 0 \\ 17 \end{pmatrix}.$$

3.4. Замена базиса

Мы на время оставим линейные преобразования и займемся подробнее заменами базиса.

Разберем фантастический пример.

Пример 2.

Рассмотрим простейший опросник, состоящий из двух вопросов:

1) Сколько у Вас друзей?

2) Какова Ваша средняя оценка по информатике?

Каждый испытуемый отвечает двумя числами, которые мы представим как вектор-столбец. Факторный анализ выявил, что результаты тестирования объясняются двумя факторами:

$\mathbf{f}_1$) уровень общего развития;

$\mathbf{f}_2$) уровень аутизации.

Если считать $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, а $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ (т.е. базисные векторы изображают вопросы опросника), то факторы представляют собой новый базис и выражаются следующими формулами

$$\mathbf{f}_1 = +0{,}71\mathbf{e}_1 + 0{,}71\mathbf{e}_2$$

$$\mathbf{f}_2 = -0{,}71\mathbf{e}_1 + 0{,}71\mathbf{e}_2.$$

Ответы на вопросы предсказываются из значений координат по факторам по формуле

$$x_e = 0{,}71x_f - 0{,}71y_f$$

$$y_e = 0{,}71x_f + 0{,}71y_f,$$

где x_e и y_e ответы на первый и второй вопросы, а x_f и y_f координаты по первому и второму факторам.

Эти формулы интерпретируются так:

— чем больше уровень развития респондента x_f, тем больше у него друзей x_e и выше оценка по информатике y_e;

— чем выше степень аутизации испытуемого, тем меньше друзей x_e, но тем выше оценка y_e.

Таким образом, некоторый аспект характера респондента описывается положением точки на плоскости. Психологи различают психологическую первичность оценок по факторам x_f и y_f и наблюдаемость признаков x_e и y_e, но для линейной алгебры это различение не важно — она занимается только формулами перехода от одного базиса к другому.

Запишем в общем случае выражение новых базисных векторов через старые:

$$\mathbf{f}_1 = a\mathbf{e}_1 + c\mathbf{e}_2$$

$$\mathbf{f}_2 = b\mathbf{e}_1 + d\mathbf{e}_2$$

(смысл необычного расположения букв b и c разъяснится ниже). Пусть вектор $\mathbf{v}$ на плоскости имеет координаты x_e, y_e в старом базисе и x_f, y_f в новом.

Последнее означает, что вектор $\mathbf{v}$ выражается через базисные векторы $\mathbf{v} = x_e\mathbf{e}_1 + y_e\mathbf{e}_2$ и $\mathbf{v} = x_f\mathbf{f}_1 + y_f\mathbf{f}_2$.

Заменим в последнем выражении векторы $\mathbf{f}_1$ и $\mathbf{f}_2$ на их выражение через $\mathbf{e}_1$ и $\mathbf{e}_2$:

$$\mathbf{v} = x_f(a\mathbf{e}_1 + c\mathbf{e}_2) + y_f(b\mathbf{e}_1 + d\mathbf{e}_2),$$

затем раскроем скобки и перегруппируем слагаемые

$$x_f(a\mathbf{e}_1 + c\mathbf{e}_2) + y_f(b\mathbf{e}_1 + d\mathbf{e}_2) = ax_f\mathbf{e}_1 + by_f\mathbf{e}_1 + cx_f\mathbf{e}_2 + dy_f\mathbf{e}_2 =$$

$$= (ax_f + by_f)\mathbf{e}_1 + (cx_f + dy_f)\mathbf{e}_2.$$

Итак, с одной стороны, $\mathbf{v} = (ax_f + by_f)\mathbf{e}_1 + (cx_f + dy_f)\mathbf{e}_2$, с другой — $\mathbf{v} = x_e\mathbf{e}_1 + y_e\mathbf{e}_2$, и это два разложения одного и того же вектора в одном и том же базисе $\{\mathbf{e}_1, \mathbf{e}_2\}$. Это значит, что равны коэффициенты разложения, т.е.

$$x_e = ax_f + by_f \text{ и } y_e = cx_f + dy_f.$$

Можно записать эти равенства в матричном виде, который понадобится нам в дальнейшем

$$\begin{pmatrix} x_e \\ y_e \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix}.$$

Чтобы читателю легче было запомнить, выражение нового базиса через старый запишем так:

$$\begin{matrix} \mathbf{f}_1 & \mathbf{f}_2 \\ \| & \| \\ a\mathbf{e}_1 & b\mathbf{e}_1 \\ + & + \\ c\mathbf{e}_2 & d\mathbf{e}_2. \end{matrix}$$

Таким образом, одна и та же матрица выражает векторы (внимание!) нового базиса через векторы старого (при вертикальном расположении равенств) и координаты всякого вектора (внимание!) в старом базисе через его же координаты в новом.

Эта матрица называется матрицей перехода от старого базиса $\{\mathbf{e}_1, \mathbf{e}_2\}$ к новому $\{\mathbf{f}_1, \mathbf{f}_2\}$.

3.5. Произведение матриц. Единичная матрица

Матрицу можно умножать не только на вектор, но и на матрицу. Произведение двух квадратных матриц 2×2 рассчитывается по следующему правилу:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix}.$$

Если мысленно разбить матрицу — второй сомножитель на два столбца, то можно привычным образом умножить левую матрицу на каждый из столбцов, а затем приложить их один к другому:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \left(\begin{array}{c|c} e & f \\ g & h \end{array}\right) = \left(\begin{array}{c|c} ae + bg & af + bh \\ ce + dg & cf + dh \end{array}\right).$$

Этот способ легко обобщается на матрицы 3×3 и выше.

Особое место среди матриц занимает так называемая единичная матрица. Она имеет единицы на главной диагонали и нули вне ее. Любая матрица, умноженная на единичную, остается без изменения:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}; \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} e & f \\ g & h \end{pmatrix}.$$

3.6. Обратная матрица

Совершенно естественно поставить вопрос о выражении координат вектора в новом базисе через его координаты в старом. Этому служит обратная матрица, произведение которой на исходную равно единичной матрице.

Обратной к матрице $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ называется такая матрица $\begin{pmatrix} e & f \\ g & h \end{pmatrix}$, что

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} e & f \\ g & h \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Матрица $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ называется единичной матрицей.

В общем случае матриц 2×2 обратной к $\begin{pmatrix} a\, b \\ c\, d \end{pmatrix}$ будет $\begin{pmatrix} d/D & -b/D \\ -c/D & a/D \end{pmatrix}$, где $D = ad - bc$ — определитель исходной матрицы. Условие $ad - bc \neq 0$ необходимо и достаточно для существования обратной матрицы.

Пусть в некоторой ситуации, похожей на разобранный выше пример опросника, координаты в базисе факторов выражались через тестовые показатели с помощью матрицы перехода

$$\begin{pmatrix} x_e \\ y_e \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x_f \\ y_f \end{pmatrix}.$$

Посчитаем обратную матрицу. Сначала вычислим определитель $D = 1 * 1 + 1 * 1 = 2$. Подставив в формулу для обратной матрицы, получим матрицу

$$\begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

Упражнение 3.3. Проверить, что перемножение этих матриц в любом порядке дает единичную матрицу.

Эта матрица связывает новые координаты и старые формулой

$$\begin{pmatrix} x_f \\ y_f \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix}\begin{pmatrix} x_e \\ y_e \end{pmatrix}.$$

Отметим, что если умножить обе части последнего равенства слева на исходную матрицу

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} x_e \\ y_e \end{pmatrix},$$

то справа произведение матриц даст единичную, которая при умножении на любой вектор оставляет его без изменения

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} x_e \\ y_e \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_e \\ y_e \end{pmatrix} = \begin{pmatrix} x_e \\ y_e \end{pmatrix},$$

и мы получаем исходное выражение старых координат через новые.

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix} = \begin{pmatrix} x_e \\ y_e \end{pmatrix}.$$

В следующей главе мы подробнее рассмотрим алгебраические операции над матрицами, а пока отметим все-таки, что этот факт является частным случаем общего тождества, связывающего любые векторы и матрицы: если для векторов-столбцов u и v и матрицы A выполнено $u = Av$, а матрица A^{-1} обратная к A, то

$$A^{-1}u = A^{-1}Av = Ev = v,$$

где E — единичная матрица.

Таким образом из

$$u = Av$$

следует

$$A^{-1}u = v.$$

3.7. Матрица линейного преобразования в новом базисе

Обычно линейное преобразование задается матрицей в некотором исходном базисе. Нам потребуется найти, как будет выглядеть это же преобразование, выраженное матрицей в другом базисе.

Рассмотрим матрицу преобразования

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

в базисе $\{\mathbf{e}_1, \mathbf{e}_2\}$ предыдущего параграфа и найдем, как это преобразование будет выглядеть в базисе $\{\mathbf{f}_1, \mathbf{f}_2\}$, связанного с исходном базисом матрицей перехода

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

Напомним, что преобразование действует на направленные отрезки на плоскости, независимо от того или иного базиса. В нашем базисе $\{\mathbf{e}_1, \mathbf{e}_2\}$ это преобразование записывается так:

$$\begin{pmatrix} x'_e \\ y'_e \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_e \\ y_e \end{pmatrix}.$$

Заметим, что

$$\begin{pmatrix} x'_f \\ y'_f \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix}$$

это то же самое преобразование, действующее на тот же самый вектор с тем же самым результатом, но только имена вектора-прообраза и вектора-образа даны в новом базисе $\{\mathbf{f}_1, \mathbf{f}_2\}$, поэтому матрица, связывающая эти новые имена, нам пока неизвестна. Найти ее совсем нетрудно.

Подставим в первую из этих двух формул преобразования вместо $\begin{pmatrix} x_e \\ y_e \end{pmatrix}$ и $\begin{pmatrix} x'_e \\ y'_e \end{pmatrix}$ их выражения через новые координаты и матрицу перехода $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix}$ и $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x'_f \\ y'_f \end{pmatrix}$ соответственно.

Получим матричное равенство

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x'_f \\ y'_f \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix}.$$

Теперь домножим обе его части слева на матрицу, обратную к матрице перехода, которую мы нашли в предыдущем параграфе,

$$\begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x'_f \\ y'_f \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix}.$$

Поскольку

$$\begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

окончательно искомая формула принимает вид

$$\begin{pmatrix} x'_f \\ y'_f \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix},$$

и нам остается только последовательно произвести перемножение матриц, чтобы найти новую матрицу старого преобразования,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1/2 & 0 \\ -1/2 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

Итак, два выражения одного и того же процесса преобразования векторов в старом и новом базисах выглядят так:

$$\begin{pmatrix} x'_e \\ y'_e \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_e \\ y_e \end{pmatrix}$$

и

$$\begin{pmatrix} x'_f \\ y'_f \end{pmatrix} = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \end{pmatrix}.$$

Если заменить матрицы и векторы их буквенными обозначениями, то предыдущие длинные выкладки запишутся в несколько строк. Преобразование $\mathbf{A}$ переводит всякий вектор $\mathbf{v}$ в соответствующий ему $\mathbf{v}'$.

В первом базисе вектор $\mathbf{v}$ имеет имя v_e, преобразование $\mathbf{A}$ имеет матрицу A_e, вектор $\mathbf{v}'$ имеет имя v'_e, а связь между ними выражается формулой

$$v'_e = A_e v_e.$$

Во втором базисе вектор $\mathbf{v}$ имеет имя v_f, преобразование имеет матрицу A_f, вектор $\mathbf{v}'$ имеет имя v'_f, а связь между ними выражается формулой

$$v'_f = A_f v_f.$$

Имена векторов в разных базисах связаны матрицей перехода, которую мы обозначим C:

$$v_e = C v_f,$$

$$v'_e = C v'_f.$$

Эти выражения подставляются в формулу

$$v'_e = A_e v_e$$

вместо v'_e и v_e. Получается формула

$$Cv'_f = A_e C v_f.$$

Домножая обе части этого равенства на матрицу C^{-1}, получаем

$$C^{-1}Cv'_f = C^{-1}A_e C v_f.$$

Но $C^{-1}Cv'_f = Ev'_f = v'_f$ (где E — единичная матрица), поэтому окончательно

$$v'_f = C^{-1}A_e C v_f.$$

Сравнивая с формулой $v'_f = A_f v_f$, заключаем, что
$C^{-1}A_e C$ и есть искомая матрица A_f:

$$A_f = C^{-1}A_e C.$$

3.8. Матрица преобразования в базисе из собственных векторов

Рассмотрим преобразование **A** векторов на плоскости (как направленных отрезков), заданное в базисе $\{\mathbf{e}_1, \mathbf{e}_2\}$ матрицей $A_e = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$.

Найдем собственные векторы этой матрицы. Для этого мы должны вычислить определитель, задающий характеристическое уравнение.

$$\begin{vmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1 = \lambda^2 - 6\lambda + 8 = 0.$$

Корни уравнения $\lambda_1 = 2$ и $\lambda_2 = 4$.

Найдем собственный вектор, соответствующий первому корню. Для этого мы должны решить систему уравнений (фактически — одно уравнение)

$$\begin{cases} (3-2)x_e & + & y_e & = & 0 \\ x_e & + & (3-2)y_e & = & 0. \end{cases}$$

Положив $y_e = 1$, получаем $x_e = -1$. Таким образом, вектор $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ является для нашей матрицы собственным вектором с собственным значением 2.

Для второго корня $\lambda = 4$ имеем систему

$$\begin{cases} (3-4)x_e + y_e = 0 \\ x_e + (3-4)y_e = 0, \end{cases}$$

эквивалентную единственному уравнению

$$-x_e + y_e = 0,$$

откуда, положив $y_e = 1$, получаем $x_e = 1$. Это значит, что вектор $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ является для нашей матрицы собственным вектором с собственным значением 4.

Но имя $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ имеет вектор $\mathbf{f_1} = -\mathbf{e_1} + \mathbf{e_2}$, а имя $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ — вектор $\mathbf{f_2} = \mathbf{e_1} + \mathbf{e_2}$.

Будем считать $\{\mathbf{f_1}, \mathbf{f_2}\}$ новым базисом для тех же векторов на плоскости.

Запишем формулы, выражающие новый базис через старый, вертикально, чтобы удобнее было определить матрицу перехода,

$$\begin{matrix} \mathbf{f_1} & \mathbf{f_2} \\ \| & \| \\ -\mathbf{e_1} & \mathbf{e_1} \\ + & + \\ \mathbf{e_2} & \mathbf{e_2}. \end{matrix}$$

Мы можем указать теперь матрицу перехода и по формуле параграфа 3.6 матрицу к ней обратную:

$$C = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}; \qquad C^{-1} = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

Упражнение 3.4. Проверить, что $CC^{-1} = E$.

Теперь вычислим матрицу того же преобразования **A** в базисе $\{f_1, f_2\}$.

$$A_f = C^{-1}A_eC = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} -1 & 1 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

В старом базисе $\{\mathbf{e}_1, \mathbf{e}_2\}$ вектор $\mathbf{f}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ был собственным с собственным значением 2, что можно проверить умножив соответствующие этому базису матрицу и вектор:

$$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}.$$

В новом базисе тот же вектор сам принадлежит базису и имеет имя $f_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Нет ничего удивительного в том, что умножение на матрицу A_f преобразования **A** в новом базисе

$$\begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix},$$

также показывает, что вектор $\mathbf{f}_1$ собственный с тем же собственным значением 2.

Эта согласованность — следствие того факта, что собственным вектором обладает само линейное преобразование **A**, действующее на векторы плоскости. Мы можем дать такое определение:

Вектор $\mathbf{v}$ называется собственным вектором линейного преобразования **A**, если $\mathbf{A}(\mathbf{v}) = \lambda \mathbf{v}$ для некоторого числа λ. Всякое выражение этого факта для конкретного базиса $\{\mathbf{g}_1, \mathbf{g}_2\}$ будет выглядеть как матричное равенство

$$A_g v_g = \lambda v_g,$$

которое будет верным в любом базисе.

Упражнение 3.5. Проверить, что вектор $\mathbf{f}_2$ является собственным для матриц преобразования **A** в обоих базисах.

В базисе из собственных векторов матрица преобразования имеет наиболее простой, так называемый диагональный вид. В общем случае это значит, что только на главной диагонали матрицы стоят ненулевые элементы (а именно собственные значения), все остальные элементы матрицы — нули. В нашем примере в базисе $\{\mathbf{f}_1, \mathbf{f}_2\}$ матрица имеет диагональный вид

$$\begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

Замечание. Далеко не все преобразования имеют достаточное количество собственных векторов, чтобы из них можно было составить базис. Совсем нет собственных векторов, например, у поворотов плоскости.

К счастью, линейные преобразования, интересующие психологов, всегда имеют базис из собственных векторов.

Глава 4

Линейные пространства, базисы, линейные преобразования

4.1. Линейные пространства

В предыдущей главе мы рассматривали пространство векторов на плоскости. Эти векторы считались чем-то реально существующим, а для удобства оперирования с ними мы давали им имена, представляющие собой наборы чисел. Мы сохраним и в данной главе представление о том, что векторы — это какие-то предметы в мире, а наборы чисел ‘— их возможные имена.

➽ **Определение 4.1.** *Линейным пространством называется множество* $\mathbf{V}$*, элементы которого называются векторами и обладают следующими свойствами:*

1. *Векторы можно складывать, т.е., если* $\mathbf{u}$ *и* $\mathbf{v}$ *векторы пространства* $\mathbf{V}$, *то и* $\mathbf{u}+\mathbf{v}$ *также элемент* $\mathbf{V}$. *Сложение коммутативно, т.е.*

$$\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u},$$

и ассоциативно, т.е.

$$\mathbf{u}+(\mathbf{v}+\mathbf{w})=(\mathbf{u}+\mathbf{v})+\mathbf{w}.$$

Существует нулевой вектор $\mathbf{0}$*, прибавление которого действует "нулевым" образом: для любого вектора* $\mathbf{v}$

$$\mathbf{0} + \mathbf{v} = \mathbf{v} + \mathbf{0} = \mathbf{0}.$$

Для каждого вектора $\mathbf{u}$ *существует противоположный ему* $\mathbf{v}$*, такой вектор, что*

$$\mathbf{u} + \mathbf{v} = \mathbf{0}.$$

Противоположный к вектору $\mathbf{u}$ *обычно обозначается* $-\mathbf{u}$.

2. *Векторы можно умножать на действительные числа, т.е., если* $\mathbf{u}$ *вектор из пространства* $\mathbf{V}$*, а* λ *действительное число, то и* $\lambda\mathbf{u}$ *также вектор из* $\mathbf{V}$*. Если* μ *также действительное число, то*

$$(\lambda\mu)\mathbf{u} = \lambda(\mu\mathbf{u}).$$

Умножение на единицу оставляет вектор без изменения

$$1\mathbf{u} = \mathbf{u}.$$

3. *Сложение векторов и умножение их на числа подчиняются естественным требованиям:*

$$(\lambda + \mu)\mathbf{u} = \lambda\mathbf{u} + \mu\mathbf{u},$$

$$\lambda(\mathbf{u} + \mathbf{v}) = \lambda\mathbf{u} + \lambda\mathbf{v}.$$

Определение линейного пространства окончено.

Исходя из перечисленных в определении свойств можно доказать, что произведение $0\mathbf{v}$ равно нулевому вектору для любого вектора $\mathbf{v}$. Поскольку определение линейного пространства дано чисто алгебраическими средствами, доказательство также чисто алгебраическое.

Во-первых, $0\mathbf{v} = (0+0)\mathbf{v} = 0\mathbf{v}+0\mathbf{v}$. Далее, к обеим частям равенства $0\mathbf{v} + 0\mathbf{v} = 0\mathbf{v}$ прибавим вектор $-0\mathbf{v}$ (противоположный к элементу $0\mathbf{v}$), в результате чего получим $0\mathbf{v} = \mathbf{0}$, что и требовалось.

Чтобы легче было различать числа и векторы, мы обозначали векторы полужирными буквами, а числа простыми. Эти особенности мы сохраним и в дальнейшем.

Исходя из определения, которое требует только возможности складывать элементы и умножать их на числа, о пространстве мало что

можно сказать. Далеко не самый странный пример линейного пространства — множество всех непрерывных функций. Их легко можно складывать и умножать на числа, причем эти операции будут обладать всеми требуемыми свойствами.

Нас будут интересовать более простые, так называемые конечномерные линейные пространства, в которых, как мы убедимся, векторы можно именовать наборами чисел некоторой фиксированной длины. Для того чтобы уметь отличать конечномерные пространства от бесконечномерных, мы введем ряд очень важных понятий.

Пусть $\mathbf{v}_1, \ldots, \mathbf{v}_k$ — векторы линейного пространства $\mathbf{V}$, а $\lambda_1, \ldots, \ldots, \lambda_k$ — действительные числа.

Выражение $\lambda_1\mathbf{v}_1 + \ldots + \lambda_k\mathbf{v}_k$ называется линейной комбинацией векторов $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Линейная комбинация называется ненулевой, если хотя бы один числовой коэффициент из $\lambda_1, \ldots, \lambda_k$ не равен нулю.

➽ **Определение 4.2.** *Векторы $\mathbf{v}_1, \ldots, \mathbf{v}_k$ называются линейно зависимыми, если существует ненулевая линейная комбинация, равная нулевому вектору:*

$$\lambda_1\mathbf{v}_1 + \ldots + \lambda_k\mathbf{v}_k = \mathbf{0}.$$

(Далее мы позволим себе говорить вместо "равная нулевому вектору" просто "равная нулю".) Если векторы не являются линейно зависимыми, то они называются линейно независимыми.

На плоскости любые три вектора линейно зависимы. Для доказательства этого факта надо разложить один из векторов по правилу параллелограмма по базису, заданному оставшейся парой векторов.

Внимание! Вышеприведенное рассуждение по математическим стандартам не годится в качестве доказательства.

Теорема 4.1. *Любые три вектора на плоскости линейно зависимы.*

Доказательство. Если хотя бы один из векторов нулевой, в линейной комбинации поставим перед ним коэффициент 1, а остальные коэффициенты сделаем нулями. Эта линейная комбинация с одним ненулевым коэффициентом будет равна нулю.

Если все три вектора отличны от нуля, применим приведенное выше и забракованное рассуждение: если какие-то два из них не лежат на одной прямой, объявим их базисом и разложим по этому базису третий вектор; если все три вектора лежат на одной прямой, то, взяв любые два из них, предположим $\mathbf{u}$ и $\mathbf{v}$, запишем $\mathbf{u} = \lambda\mathbf{v}$, откуда $\mathbf{u} - \lambda\mathbf{v} = \mathbf{0}$.

Мы можем теперь определить конечномерное пространство.

Определение 4.3. *Линейное пространство имеет размерность меньше n, если любые n его векторов линейно зависимы.*

Пространство векторов на плоскости имеет размерность меньше 140, 62, 5 и 3.

Определение 4.4. *Линейное пространство имеет размерность n, если его размерность меньше $n+1$ и существуют n линейно независимых векторов этого пространства.*

Пространство векторов на плоскости имеет размерность 2.

Линейное пространство размерности n мы будем коротко называть n-мерным линейным пространством.

Теорема 4.2. *Если $\mathbf{v_1}, \ldots, \mathbf{v_n}$ — n линейно независимых векторов в n-мерном пространстве, то любой вектор этого пространства выражается через них, или, другими словами, является их линейной комбинацией.*

Доказательство. Пусть $\mathbf{w}$ — вектор из нашего пространства. Поскольку размерность пространства меньше $n+1$, т.е. любые $n+1$ векторов линейно зависимы, то линейно зависимы векторы $\mathbf{w}, \mathbf{v_1}, \ldots, \mathbf{v_n}$. Это значит, что некоторая их ненулевая линейная комбинация равна нулю:

$$\mu\mathbf{w} + \lambda_1\mathbf{v_1} + \ldots + \lambda_n\mathbf{v_n} = \mathbf{0}.$$

Если бы оказалось, что в этой комбинации $\mu = 0$, то из этого следовало бы, что

$$\lambda_1\mathbf{v_1} + \ldots + \lambda_n\mathbf{v_n} = \mathbf{0}$$

(причем один из коэффициентов отличен от нуля — иначе не была бы ненулевой линейная комбинация $\mu\mathbf{w} + \lambda_1\mathbf{v_1} + \ldots + \lambda_n\mathbf{v_n} = \mathbf{0}$). Этого, однако, не может быть, поскольку векторы $\mathbf{v_1}, \ldots \mathbf{v_n}$ линейно независимы. Это значит, что $\mu \neq 0$.

В таком случае из

$$\mu\mathbf{w} + \lambda_1\mathbf{v_1} + \ldots + \lambda_n\mathbf{v_n} = \mathbf{0}$$

следует

$$\mathbf{w} = -\lambda_1/\mu\mathbf{v_1} - \ldots - \lambda_n/\mu\mathbf{v_n}.$$

Теорема доказана.

Мы будем назвать систему векторов $\mathbf{v_1}, \ldots, \mathbf{v_n}$ базисом n-мерного пространства. Разложив вектор по базису, мы можем именовать его соответствующим столбцом чисел. В этом случае, если

$$\mathbf{w} = x_1\mathbf{v_1} + \ldots + x_n\mathbf{v_n},$$

то

$$w_v = \begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix}.$$

Открытым остается, правда, вопрос, не может ли какая-то другая система, содержащая меньше векторов (например $n-1$), также быть базисом в этом пространстве? Можно доказать, что базис в n-мерном пространстве всегда содержит ровно n векторов, но это доказательство мы здесь приводить не будем.

4.2. Линейные преобразования. Связь преобразования, базиса и матрицы

Определение линейного преобразования, данное в предыдущей главе, без изменений приложимо к любому линейному пространству. Повторим его коротко.

Линейное преобразование $\mathbf{A}$ ставит в соответствие каждому вектору $\mathbf{u}$ некоторый другой вектор $\mathbf{u}' = \mathbf{A}(\mathbf{u})$, причем

$$\mathbf{A}(\lambda\mathbf{u} + \mu\mathbf{v}) = \lambda\mathbf{A}(\mathbf{u}) + \mu\mathbf{A}(\mathbf{v}) \tag{4.1}$$

Пусть $\mathbf{e_1}, \ldots, \mathbf{e_n}$ — базис в пространстве $\mathbf{V}$, а $\mathbf{A}$ — линейное преобразование этого пространства.

Если некоторый вектор имеет в нашем базисе следующее разложение

$$\mathbf{w} = w_1\mathbf{e_1} + \ldots + w_n\mathbf{e_n},$$

то по формуле (4.1) его образ выразится через образы базисных векторов:

$$\mathbf{u} = \mathbf{A}(\mathbf{w}) = w_1\mathbf{A}(\mathbf{e_1}) + \ldots + w_n\mathbf{A}(\mathbf{e_n}). \tag{4.2}$$

Разложим каждый из векторов-образов $\mathbf{A}(\mathbf{e_1}), \ldots, \mathbf{A}(\mathbf{e_n})$ по нашему базису и поставим в соответствие каждому из них вектор-столбец:

$$\mathbf{u} = \mathbf{A}(\mathbf{w}) \Rightarrow \begin{pmatrix} u_1 \\ \dots \\ u_n \end{pmatrix}, \mathbf{A}(\mathbf{e_1}) \Rightarrow \begin{pmatrix} a_{11} \\ \dots \\ a_{n1} \end{pmatrix}, \dots, \mathbf{A}(\mathbf{e_n}) \Rightarrow \begin{pmatrix} a_{n1} \\ \dots \\ a_{nn} \end{pmatrix}.$$

Упражнение 4.1. Доказать, что разложение по базису суммы векторов равно сумме их разложений и что разложение данного ветора по данному базису единственно.

Теперь заменим в равенстве (4.2) векторы на столбцы, представляющие их разложения по базису

$$\begin{pmatrix} u_1 \\ \dots \\ u_n \end{pmatrix} = w_1 \begin{pmatrix} a_{11} \\ \dots \\ a_{n1} \end{pmatrix} + \dots + w_n \begin{pmatrix} a_{1n} \\ \dots \\ a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11}w_1 + a_{12}w_2 + \dots + a_{1n}w_n \\ \dots \\ a_{n1}w_1 + a_{n2}w_2 + \dots + a_{nn}w_n \end{pmatrix}.$$

Справа стоит вектор, который можно получить перемножением матрицы и столбца:

$$\begin{pmatrix} u_1 \\ \dots \\ u_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & & \dots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} w_1 \\ \dots \\ w_n \end{pmatrix}.$$

Это равенство получено для произвольного вектора $\mathbf{w}$, следовательно, мы показали, что действие нашего преобразования на векторы столбцы, именующие векторы пространства $\mathbf{V}$ в базисе $\{\mathbf{e_1}, \dots, \mathbf{e_n}\}$, описывается умножением на матрицу

$$A_e = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & & \dots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}.$$

Коротко: если для векторов пространства $\mathbf{V}$ и оператора $\mathbf{A}$ выполнено

$$\mathbf{u} = \mathbf{A}(\mathbf{w}),$$

то для их разложений по базису и соответствующей матрицы

$$u_e = A_e w_e.$$

4.3. Замена базиса. Матрица перехода

Как и в предыдущей главе, запишем выражение нового базиса $\{\mathbf{f_1},\ldots,\mathbf{f_n}\}$ через старый $\{\mathbf{e_1},\ldots,\mathbf{e_n}\}$ вертикальными равенствами:

$$\begin{array}{cccc}
\mathbf{f_1} & \mathbf{f_2} & \ldots & \mathbf{f_n} \\
\| & \| & & \| \\
c_{11}\mathbf{e_1} & c_{12}\mathbf{e_1} & \ldots & c_{1n}\mathbf{e_1} \\
+ & + & & + \\
c_{21}\mathbf{e_2} & c_{22}\mathbf{e_2} & \ldots & c_{2n}\mathbf{e_2} \\
+ & + & & + \\
\ldots & \ldots & \ldots & \ldots \\
+ & + & & + \\
c_{n1}\mathbf{e_n} & c_{n2}\mathbf{e_n} & \ldots & c_{nn}\mathbf{e_n}.
\end{array} \tag{4.3}$$

Пусть произвольный вектор $\mathbf{w}$ разлагается в новом базисе в сумму: $\mathbf{w} = w'_1\mathbf{f_1}+\ldots+w'_n\mathbf{f_n}$. Подставим в разложение вместо $\mathbf{f_i}$ их выражения через векторы старого базиса и представим результат в виде таблицы:

$$\begin{array}{ccccccccc}
\mathbf{w} & = & w'_1\mathbf{f_1} & + & w'_2\mathbf{f_2} & + & \ldots & + & w'_n\mathbf{f_n}
\end{array}$$

$$\begin{array}{ccccccccc}
 & & w'_1c_{11}\mathbf{e_1} & + & w'_2c_{12}\mathbf{e_1} & + & \ldots & + & w'_nc_{1n}\mathbf{e_1} \\
 & & + & & + & & & & + \\
 & & w'_1c_{21}\mathbf{e_2} & + & w'_2c_{22}\mathbf{e_2} & + & \ldots & + & w'_nc_{2n}\mathbf{e_2} \\
\mathbf{w} & = & + & & + & & & & + \\
 & & \ldots & & \ldots & & \ldots & & \ldots \\
 & & + & & + & & & & + \\
 & & w'_1c_{n1}\mathbf{e_n} & + & w'_2c_{n2}\mathbf{e_n} & + & \ldots & + & w'_nc_{nn}\mathbf{e_n}.
\end{array}$$

В верхней строке стоит разложение вектора $\mathbf{w}$ по новому базису, и под каждым элементом верхней суммы помещен результат подстановки вместо $\mathbf{f_i}$ его разложения в формулах, помеченных (4.3).

В каждой строке таблицы стоят кратные одного и того же вектора старого базиса. Вынесем $\mathbf{e_i}$ за скобки и получим

$$\mathbf{w} =$$

$$= (w'_1c_{11} + w'_2c_{12} + \ldots + w'_nc_{1n})\mathbf{e_1}+$$

$$+(w'_1c_{21} + w'_2c_{22} + \ldots + w'_nc_{2n})\mathbf{e_2}+$$

$$+\ldots+$$

$$+(w'_1 c_{n1} + w'_2 c_{n2} + \ldots + w'_n c_{nn})\mathbf{e_n}.$$

Это и есть разложение $\mathbf{w}$ по старому базису. Если обозначить координаты $\mathbf{w}$ в старом базисе через $w_1, \ldots, w_n$, то старые и новые координаты связывает система равенств

$$w_1 = w'_1 c_{11} + w'_2 c_{12} + \ldots + w'_n c_{1n}$$

$$w_2 = w'_1 c_{21} + w'_2 c_{22} + \ldots + w'_n c_{2n}$$

$$\ldots$$

$$w_n = w'_1 c_{n1} + w'_2 c_{n2} + \ldots + w'_n c_{nn}.$$

Эту систему можно заменить на матричное равенство

$$\begin{pmatrix} w_1 \\ w_2 \\ \ldots \\ w_n \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{pmatrix} \begin{pmatrix} w'_1 \\ w'_2 \\ \ldots \\ w'_n \end{pmatrix}$$

или, еще короче, в виде соотношения $w_e = Cw_f$.

4.4. Произведение матриц. Единичная матрица

В ближайших параграфах нам понадобятся некоторые сведения из алгебры матриц. Мы сейчас введем только одну операцию — умножение, остальные сведения можно найти в начале восьмой главы.

➽ **Определение 4.5.** *Две матрицы можно перемножить, если их размеры соответствуют друг другу. Если умножается матрица $k \times n$ на матрицу $m \times k$, то результатом будет матрица размера $m \times n$. Правило умножения таково:*

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nk} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \ldots & \ldots & b_{1m} \\ b_{21} & b_{22} & \ldots & \ldots & b_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{k1} & b_{k2} & \ldots & \ldots & b_{km} \end{pmatrix} =$$

$$= \begin{pmatrix} c_{11} & c_{12} & \ldots & \ldots & c_{1m} \\ c_{21} & c_{22} & \ldots & \ldots & c_{2m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{n1} & c_{n2} & \ldots & \ldots & c_{nm} \end{pmatrix},$$

где $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{ik}b_{kj}$.

Правило поясним следующей схемой:

$$\begin{pmatrix} \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ a_{i1} & a_{i2} & \ldots & a_{ik} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix} \begin{pmatrix} \ldots & b_{1j} & \ldots \\ \ldots & b_{2j} & \ldots \\ \ldots & \ldots & \ldots \\ \ldots & b_{kj} & \ldots \end{pmatrix} = \begin{pmatrix} \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots \\ \ldots & c_{ij} & \ldots \\ \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots \end{pmatrix}.$$

Обратим внимание на две вещи.

1) Чтобы подсчитать c_{ij}, надо первый элемент выделенной строки умножить на первый элемент выделенного столбца, к этому прибавить произведение второго элемента выделенной строки на второй элемент выделенного столбца, ..., прибавить произведение k-го (последнего) элемента выделенной строки на k-й (последний) элемент выделенного столбца. Если ширина первой матрицы не равна высоте второй, то перемножение данных матриц невозможно.

2) Матрица-результат наследует высоту первой матрицы и ширину второй.

➽ **Определение 4.6.** *Единичной матрицей порядка n будем называть матрицу размера $n \times n$*

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

с единицами на главной диагонали и нулями вне ее. Единичная матрица любого порядка обозначается одинаково — знаком E. Размер единичной матрицы всегда ясен из контекста рассуждения.

Умножение единичной матрицы на любой вектор-столбец оставляет этот столбец без изменений:

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}.$$

Аналогично и произведение любой матрицы и единичной оставляет матрицу без изменения:

$$AE = EA = A.$$

Упражнение 4.2. Проверить последнее тождество для матриц 4×4.

4.5. Обратная матрица

В параграфе 2.3 главы 2 было введено понятие алгебраического дополнения элемента квадратной матрицы. Чтобы подсчитать алгебраическое дополнение (напомним, что оно обозначается $\overline{\overline{A}}_{ij}$) элемента a_{ij} в матрице A, надо вычеркнуть в матрице i-ю строку и j-й столбец, подсчитать определитель оставшейся матрицы и умножить его на $(-1)^{i+j}$.

➽ **Определение 4.7.** *(Обобщение определения 3 параграфа 2.3)* Определитель матрицы может быть вычислен не только по формуле

$$\det A = a_{11}\overline{\overline{A}}_{11} + a_{12}\overline{\overline{A}}_{12} + \ldots + a_{1n}\overline{\overline{A}}_{1n},$$

но и по аналогичной формуле, связывающей элементы строки и их алгебраические дополнения:

$$\det A = a_{i1}\overline{\overline{A}}_{i1} + a_{i2}\overline{\overline{A}}_{i2} + \ldots + a_{in}\overline{\overline{A}}_{in},$$

а также и по формуле

$$\det A = a_{1i}\overline{\overline{A}}_{1i} + a_{2i}\overline{\overline{A}}_{2i} + \ldots + a_{ni}\overline{\overline{A}}_{ni},$$

связывающей элементы произвольного столбца и их алгебраические дополнения.

Упражнение 4.3. Доказать первую формулу перестановкой строк, а вторую транспонированием.

➽ **Определение 4.8.** *Обратной к матрице A называется матрица B, такая, что $AB = BA = E$. Обратная матрица обычно обозначается A^{-1}.*

Обратная матрица к матрице A расчитывается по формуле

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} \overline{\overline{A}}_{11} & \overline{\overline{A}}_{21} & \ldots & \overline{\overline{A}}_{n1} \\ \overline{\overline{A}}_{12} & \overline{\overline{A}}_{22} & \ldots & \overline{\overline{A}}_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ \overline{\overline{A}}_{1n} & \overline{\overline{A}}_{2n} & \ldots & \overline{\overline{A}}_{nn} \end{pmatrix}.$$

(Обратите внимание на индексы! Их порядок необычный.)

Знак $\frac{1}{\det A}$ перед матрицей означает, что все элементы матрицы надо поделить на $\det A$, подробнее об алгебре матриц в главе 8.

Легко проверить, что это действительно обратная матрица, т.е.

$$\frac{1}{\det A}\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}\begin{pmatrix} \overline{\overline{A}}_{11} & \overline{\overline{A}}_{21} & \dots & \overline{\overline{A}}_{n1} \\ \overline{\overline{A}}_{12} & \overline{\overline{A}}_{22} & \dots & \overline{\overline{A}}_{n2} \\ \dots & \dots & \dots & \dots \\ \overline{\overline{A}}_{1n} & \overline{\overline{A}}_{2n} & \dots & \overline{\overline{A}}_{nn} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Действительно, в верхнем левом углу матрицы-произведения окажется число

$$\frac{1}{\det A}(a_{11}\overline{\overline{A}}_{11} + a_{12}\overline{\overline{A}}_{12} + \dots a_{1n}\overline{\overline{A}}_{1n}).$$

Но $a_{11}\overline{\overline{A}}_{11}+a_{12}\overline{\overline{A}}_{12}+\dots a_{1n}\overline{\overline{A}}_{1n}$ в точности совпадает с определением 4.7 (даже в необобщенной форме) — это и есть $\det A$. Следовательно, в верхнем левом углу окажется единица. Для доказательства того, что и на остальных местах главной диагонали окажутся единицы, надо выписать соответствующие выражения и убедиться, что они выражают как раз $\det A$ по обобщенному определению 4.7.

Чуть труднее убедиться, что остальные элементы произведения — нули. Разберем простейший случай. Во второй строке на первом месте в матрице-произведении будет стоять

$$\frac{1}{\det A}(a_{21}\overline{\overline{A}}_{11} + a_{22}\overline{\overline{A}}_{12} + \dots a_{2n}\overline{\overline{A}}_{1n}).$$

Рассмотрим матрицу

$$\begin{pmatrix} a_{21} & a_{22} & \dots & a_{2n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix},$$

в которой две первых строки совпадают. Ее определитель в силу этого совпадения равен нулю. Но $a_{21}\overline{\overline{A}}_{11} + a_{22}\overline{\overline{A}}_{12} + \dots a_{2n}\overline{\overline{A}}_{1n}$ и есть расчет

этого определителя по формуле определения 4.7 (в общую формулу расчета определителя надо подставить a_{21} вместо a_{11}, a_{22} вместо a_{12} и т.д.).

Аналогично доказывается, что нули заполняют всю оставшуюся матрицу — для доказательства надо воспользоваться обобщенным определением 4.7.

Совершенно аналогично, используя вторую часть обобщенного определения 4.7, можно доказать, что перемножение матриц в другом порядке также дает единичную матрицу:

$$\frac{1}{\det A}\begin{pmatrix} \overline{\overline{A}}_{11} & \overline{\overline{A}}_{21} & \dots & \overline{\overline{A}}_{n1} \\ \overline{\overline{A}}_{12} & \overline{\overline{A}}_{22} & \dots & \overline{\overline{A}}_{n2} \\ \dots & \dots & \dots & \dots \\ \overline{\overline{A}}_{1n} & \overline{\overline{A}}_{2n} & \dots & \overline{\overline{A}}_{nn} \end{pmatrix}\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Таким образом, мы умеем вычислять обратную матрицу для всякой матрицы, у которой определитель не равен нулю. Если $\det A = 0$, множитель $1/(\det A)$ не позволяет воспользоваться нашей формулой. В главе 8 будет показано, что такая матрица не имеет обратной.

Независимо от природы объектов, скрывающихся за векторами-столбцами и матрицами, мы можем использовать обратную матрицу чисто алгебраически. В любом случае, если матрица M связывает столбцы u и v соотношением $u = Mv$, а M^{-1} обратная к M, то $M^{-1}u = v$: если домножить равенство $u = Mv$ слева на матрицу M^{-1}, получим $M^{-1}u = M^{-1}Mv = Ev = v$.

В предыдущем параграфе мы связали координаты вектора в старом и новом базисе соотношением

$$w_e = Cw_f.$$

Из сказанного выше следует, что

$$w_f = C^{-1}w_e.$$

4.6. Матрица линейного преобразования в новом базисе

Пусть линейное преобразование действует в линейном пространстве $\mathbf{V}$ и $\mathbf{u} = \mathbf{A}(\mathbf{w})$. Матричная форма связывает соответствующие столбцы и матрицу в базисе $\{\mathbf{e_i}\}$:

$$u_e = A_e w_e.$$

Поскольку $u_e = Cu_f$ и $w_e = Cw_f$, прямая подстановка в предыдущую формулу дает

$$Cu_f = A_e C w_f.$$

Домножим это алгебраическое равенство на C^{-1} слева:

$$C^{-1}Cu_f = C^{-1}A_e C w_f.$$

Поскольку $C^{-1}Cu_f = Eu_f = u_f$, то

$$u_f = C^{-1}A_e C w_f.$$

Таким образом, для любых двух векторов $\mathbf{u}$ и $\mathbf{w}$, связанных линейным преобразованием $\mathbf{u} = \mathbf{A}(\mathbf{w})$, выполнено соотношение между представляющими их в базисе $\{\mathbf{f_i}\}$ столбцами

$$u_f = C^{-1}A_e C w_f,$$

т.е. $C^{-1}A_e C$ и есть матрица A_f, представляющая линейное преобразование $\mathbf{A}$ в базисе $\{\mathbf{f_i}\}$.

Таким образом, если C — матрица перехода от старого базиса к новому, а A_e — матрица линейного преобразования в старом базисе, то матрица этого же преобразования в новом базисе вычисляется по формуле

$$A_f = C^{-1}A_e C.$$

4.7. Матрица линейного преобразования в базисе из собственных векторов

Если у линейного преобразования $\mathbf{A}$, действующего в n-мерном линейном пространстве $\mathbf{V}$, имеется n линейно независимых собственных векторов, то, взяв их в качестве базиса линейного пространства $\mathbf{V}$, мы получим наиболее простой вид матрицы преобразования.

Пусть $\mathbf{v_1}, \ldots, \mathbf{v_n}$ — базис из собственных векторов, причем $\mathbf{A}(\mathbf{v_i}) = \lambda_i \mathbf{v_i}$. Обозначим A_v матрицу преобразования $\mathbf{A}$ в этом базисе. Вектор v_1 в базисе, первым элементом которого он является, имеет разложение $\begin{pmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{pmatrix}$, а чтобы получить разложение вектора-образа $\mathbf{A}(\mathbf{v_1})$ надо этот столбец умножить на матрицу A_v. Получаем

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ \ldots \\ a_{n1} \end{pmatrix}.$$

Но вектор v_1 собственный, поэтому

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \ldots \\ a_{n1} \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ 0 \\ \ldots \\ 0 \end{pmatrix}.$$

Аналогично $\begin{pmatrix} a_{12} \\ a_{22} \\ \ldots \\ a_{n2} \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda_2 \\ \ldots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} a_{1n} \\ a_{2n} \\ \ldots \\ a_{nn} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \ldots \\ \lambda_n \end{pmatrix}.$

Следовательно, матрица A_v имеет диагональный вид

$$A_v = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}.$$

Глава 5

Линейные преобразования в евклидовых пространствах. Идеи и примеры

5.1. Евклидовы пространства

Общие теоремы о замене базиса в линейном пространстве проще было доказывать, не предполагая, что векторы, образующие базис, перпендикулярны друг другу. Теперь мы вернемся к привычному для школьной геометрии понятию базиса с ортогональными векторами единичной длины.

Для измерения углов в двух- и трехмерном пространствах, на примерах которых в данной главе будет рассматриваться следующий круг понятий линейной алгебры, нам необходимо ввести понятие скалярного произведения векторов (как направленных отрезков, имеющих общее начало, — определение было дано в главе 3).

➽ **Определение 5.1.** *Скалярным произведением двух векторов* $\mathbf{u}$ *и* $\mathbf{v}$ *является число, равное произведению длин этих векторов, умноженному на косинус угла между ними.*

Для скалярного произведения вводится обозначение $(\mathbf{u}, \mathbf{v})$. Определение 5.1 выражается тогда формулой $(\mathbf{u}, \mathbf{v}) = |\mathbf{u}||\mathbf{v}|\cos\alpha$, где α — угол между векторами $\mathbf{u}$ и $\mathbf{v}$, $|\mathbf{u}|$ и $|\mathbf{v}|$ их длины.

Упражнение 5.1. Доказать, что $(\lambda\mathbf{u}, \mathbf{v}) = \lambda(\mathbf{u}, \mathbf{v})$.

➽ **Определение 5.2.** *Базис $\{\mathbf{e}_1, \mathbf{e}_2\}$ называется ортонормированным, если входящие в него векторы имеют единичную длину и взаимно перпендикулярны. Определение для трехмерного пространства точно такое же.*

Теорема 5.1. *Пусть $\mathbf{u}$ и $\mathbf{v}$ векторы, а $\begin{pmatrix} x_u \\ y_u \end{pmatrix}$ и $\begin{pmatrix} x_v \\ y_v \end{pmatrix}$ их координатное выражение в некотором ортонормированном базисе $\{\mathbf{e}_1, \mathbf{e}_2\}$.*

Тогда $(\mathbf{u}, \mathbf{v}) = x_u x_v + y_u y_v$.

Доказательство. В формуле $(\mathbf{u}, \mathbf{v}) = |\mathbf{u}||\mathbf{v}|\cos\alpha$ можно заменить $|\mathbf{v}|\cos\alpha$ на $Pr_u\mathbf{v}$ — проекцию вектора $\mathbf{v}$ на вектор $\mathbf{u}$ (рис. 5.1: $CO = Pr_u\mathbf{v}_1 = |\mathbf{v}_1|\cos\alpha$), т.е. $(\mathbf{u}, \mathbf{v}) = |\mathbf{u}|Pr_u\mathbf{v}$.

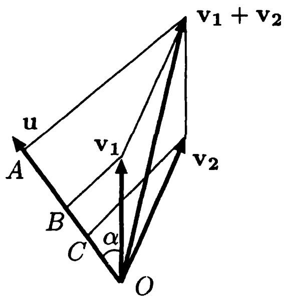


Рис. 5.1. Векторы, их сумма и соответствующие проекции

Если $\mathbf{v}_1$ и $\mathbf{v}_2$ любые два вектора, то проекция их суммы равна сумме проекций (см. рис. 5.1): поскольку $AB = CO$, то сумма проекций CO и BO векторов v_1 и v_2 равна AO — проекции вектора $\mathbf{v}_1 + \mathbf{v}_2$, следовательно $Pr_u(\mathbf{v}_1 + \mathbf{v}_2) = Pr_u\mathbf{v}_1 + Pr_u\mathbf{v}_2$.

Этого достаточно, чтобы показать, что $(\mathbf{u}, \mathbf{v}_1 + \mathbf{v}_2) = (\mathbf{u}, \mathbf{v}_1) + (\mathbf{u}, \mathbf{v}_2)$, поскольку $(\mathbf{u}, \mathbf{v}_1 + \mathbf{v}_2) = |\mathbf{u}|Pr_u(\mathbf{v}_1 + \mathbf{v}_2) = |\mathbf{u}|(Pr_u\mathbf{v}_1 + Pr_u\mathbf{v}_2) = |\mathbf{u}|Pr_u\mathbf{v}_1 + |\mathbf{u}|Pr_u\mathbf{v}_2 = (\mathbf{u}, \mathbf{v}_1) + (\mathbf{u}, \mathbf{v}_2)$.

Точно так же и $(\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}) = (\mathbf{u}_1, \mathbf{v}) + (\mathbf{u}_2, \mathbf{v})$.

Выразим теперь векторы $\mathbf{u}$ и $\mathbf{v}$ через векторы базиса и перемножим скалярно эти выражения: $\mathbf{u} = x_u\mathbf{e}_1 + y_u\mathbf{e}_2; \mathbf{v} = x_v\mathbf{e}_1 + y_v\mathbf{e}_2$, поэтому (см. упр. 5.1)

$$(\mathbf{u}, \mathbf{v}) = (x_u\mathbf{e}_1 + y_u\mathbf{e}_2, x_v\mathbf{e}_1 + y_v\mathbf{e}_2).$$

По только что доказанному, скалярное произведение сумм векторов раскладывается в сумму скалярных произведений:

$$(x_u\mathbf{e}_1 + y_u\mathbf{e}_2, x_v\mathbf{e}_1 + y_v\mathbf{e}_2) =$$

$$= x_u x_v(\mathbf{e}_1, \mathbf{e}_1) + x_u y_v(\mathbf{e}_1, \mathbf{e}_2) + y_u x_v(\mathbf{e}_2, \mathbf{e}_1) + y_u y_v(\mathbf{e}_2, \mathbf{e}_2).$$

Остается заметить, что в ортонормированном базисе

$$(\mathbf{e}_1, \mathbf{e}_1) = (\mathbf{e}_2, \mathbf{e}_2) = 1, \text{ a } (\mathbf{e}_2, \mathbf{e}_1) = (\mathbf{e}_1, \mathbf{e}_2) = \mathbf{0},$$

поэтому окончательно

$$(\mathbf{u}, \mathbf{v}) = x_u y_u + x_v y_v.$$

Следствие. Длина вектора u, представленного столбцом $\begin{pmatrix} x_u \\ y_u \end{pmatrix}$, в любом ортонормированном базисе выражается формулой $\sqrt{x_u^2 + y_u^2}$.

Поскольку косинус нулевого угла равен единице, $|\mathbf{u}| = \sqrt{|\mathbf{u}||\mathbf{u}|} = \sqrt{(\mathbf{u}, \mathbf{u})} = \sqrt{\mathrm{x}_\mathrm{u}^2 + \mathrm{y}_\mathrm{u}^2}$.

Следствие доказано.

Наиболее интересно в утверждении теоремы 5.1 то, что форма выражения скалярного произведения в виде попарного произведения координат векторов-сомножителей не зависит от базиса, хотя численные выражения координат, разумеется, от базиса зависят.

Мы будем рассматривать далее векторы на плоскости, но перенести рассуждения на трехмерный случай не составит труда. Мы советуем читателю если не проводить этот перенос самостоятельно, то по крайней мере следить за тем, что в рассуждениях будет меняться при таком переносе — обнаружится, что изменения касаются только количества слагаемых.

5.2. Замена ортонормированного базиса. Ортогональные матрицы

Здесь мы продемонстрируем пример поразительной эффективности скалярного произведения при рассмотрении замены ортонормированных базисов.

Если векторы ортогональны, то косинус угла между ними равен нулю, следовательно, равно нулю и их скалярное произведение.

Если $\{\mathbf{f}_1, \mathbf{f}_2\}$ другой ортонормированный базис на плоскости, с формулами перехода

$$\begin{array}{cc} \mathbf{f}_1 & \mathbf{f}_2 \\ \| & \| \\ a\mathbf{e}_1 & b\mathbf{e}_1 \\ + & + \\ c\mathbf{e}_2 & d\mathbf{e}_2, \end{array}$$

то векторы $\mathbf{f}_1$ и $\mathbf{f}_2$ представляются в базисе $\{\mathbf{e}_1, \mathbf{e}_2\}$ столбцами

$$\begin{pmatrix} a \\ c \end{pmatrix} \text{ и } \begin{pmatrix} b \\ d \end{pmatrix}.$$

Поскольку векторы $\mathbf{f}_1$ и $\mathbf{f}_2$ ортогональны, то $ab + cd = 0$, поскольку они имеют единичную длину, то $a^2 + c^2 = 1$ и $b^2 + d^2 = 1$.

Последние соотношения означают: даже не производя вычислений, мы можем сказать, что обратной к матрице

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ будет транспонированная матрица } \begin{pmatrix} a & c \\ b & d \end{pmatrix},$$

поскольку

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{pmatrix}.$$

Упражнение 5.2. Обобщить рассуждение на случай трех, а затем и большего числа измерений.

Матрица, у которой скалярные квадраты столбцов равны единице, а попарные скалярные произведения столбцов равны нулю, называются ортогональными.

Говоря о скалярном произведении столбцов произвольной матрицы, мы допускаем некоторую вольность. В общем случае подразумевается суммирование попарных произведений: верхнего элемента первого столбца на верхний второго, к которым прибавляется произведение второго элемента на второй и т.д. В нашем случае столбцы представляют собой разложения базисных векторов, поэтому скалярное произведение столбцов можно понимать почти строго.

Рассмотрим ортогональную в смысле нашего определения матрицу

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}.$$

Упомянутое в определении свойство столбцов эквивалентно тому, что

$$\begin{pmatrix} a & d & g \\ b & e & h \\ c & f & i \end{pmatrix} \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Вовсе не очевидно, что в таком случае и строки матрицы обладают таким же свойством: их скалярные квадраты равны единице, а попарные произведения — нулю. Даже в случае 3×3 непосредственное алгебраическое доказательство этого факта очень трудно. Однако оказывается, что доказать этот факт в самом общем случае просто, если "просто" понимать в специфическом смысле.

Есть несколько не слишком трудных, но, самое главное, весьма естественных способов доказательства того, что ортогональная матрица имеет обратную. Мы не будем их приводить, поскольку это интуитивно очевидно — обратной к матрице перехода от старого базиса к новому будет матрица обратного перехода от нового базиса к старому. Эта обратная матрица и есть транспонированная прежняя, а свойство ортогональности ее столбцов — это искомое свойство ортогональности строк матрицы перехода от старого базиса к новому.

5.3. Самосопряженные линейные преобразования

В предыдущем параграфе мы показали, что матрица перехода от ортонормированного базиса к ортонормированному является ортогональной, т.е. $C^{-1} = C'$, где C' здесь и дальше будет обозначать у нас транспонированную матрицу C.

Это значит, что матрица линейного преобразования в новом ортонормированном базисе может быть выражена через матрицу этого преобразования в старом ортонормированном базисе также и формулой $C'AC$. Этот факт имеет для нас важные последствия.

Факторный анализ имеет дело с симметричными матрицами, т.е. такими, у которых элементы a_{ij} и a_{ji}, расположенные симметрично относительно главной диагонали, равны между собой.

Мы сейчас покажем, что произведение матриц $C'AC$ всегда будет симметричным, если матрица A симметрична.

Рассмотрим простейший пример с симметричной матрицей посредине и с соответствующим отношением первой и третьей матриц:

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} f & g \\ g & h \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} af+gc & ag+ch \\ bf+dg & bg+dh \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} =$$

$$= \begin{pmatrix} \dots & (afb+gcb)+(agd+chd) \\ (abf+adg)+(bgc+dhc) & \dots \end{pmatrix}.$$

Поскольку элементы главной диагонали нас не интересуют, мы не стали их считать и заменили многоточиями. Равенство симметричных элементов имеет место, однако причины этого равенства совершенно неясны. Они становятся ясными в свете более общих соображений.

Заметим, во-первых, что $C'A$ является транспонированной к AC :

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} f & g \\ g & h \end{pmatrix} = \begin{pmatrix} af+gc & ag+ch \\ bf+dg & bg+dh \end{pmatrix};$$

$$\begin{pmatrix} f & g \\ g & h \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} af+gc & bf+dg \\ ag+ch & bg+dh \end{pmatrix}.$$

Но дальнейшие рассуждения станут еще яснее, если увидеть, что этот факт является следствием еще более общего факта: матрица MN является отражением относительно главной диагонали матрицы $N'M'$ вообще всегда. На примере 2×2 это совершенно очевидно:

$$\begin{pmatrix} k & l \\ m & n \end{pmatrix}\begin{pmatrix} o & p \\ r & s \end{pmatrix} = \begin{pmatrix} ko+lr & kp+ls \\ mo+nr & mp+ns \end{pmatrix};$$

$$\begin{pmatrix} o & r \\ p & s \end{pmatrix}\begin{pmatrix} k & m \\ l & n \end{pmatrix} = \begin{pmatrix} ok+rl & om+rn \\ pk+sl & pm+sn \end{pmatrix}.$$

(Мы советуем читателю прослеживать совпадение операций, а не результатов.) Итак, мы убедились, что $(MN)' = N'M'$ для любых матриц M и N. Отсюда следует, что $(MNO)' = O'N'M'$ для любых матриц M, N и O, поскольку $(M(NO))' = (NO)'M' = O'N'M'$.

Возвращаясь к нашей теме: по последнему равенству транспонирование матрицы $C'AC$ приведет к результату $C'A'C$, поскольку $(C')' = C$. Здесь мы и можем использовать симметричность матрицы $A = A'$, и тем самым транспонированная к $C'AC$ есть она сама, т.е. она симметрична.

Это означает, что линейное преобразование, у которого хотя бы в одном ортонормированном базисе матрица симметрична, имеет симметричную матрицу в любом ортонормированном базисе.

Такие преобразования называются самосопряженными.

Симметричность матрицы A имеет важные следствия. Первое из них состоит в том, что $(Au, v) = (u, Av)$ для любых векторов-столбцов u и v. Покажем это на примере 2×2.

$$\left(\begin{pmatrix} a & b \\ b & d \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right) = \left(\begin{pmatrix} au_1 + bu_2 \\ bu_1 + du_2 \end{pmatrix}, \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right) =$$

$$= (au_1 + bu_2)v_1 + (bu_1 + du_2)v_2 = au_1v_1 + b(u_2v_1 + u_1v_2) + du_2v_2,$$

$$\left(\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \begin{pmatrix} a & b \\ b & d \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right) = \left(\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}\begin{pmatrix} av_1 + bv_2 \\ bv_1 + dv_2 \end{pmatrix},\right) =$$

$$= u_1(av_1 + bv_2) + u_2(bv_1 + dv_2) = au_1v_1 + b(u_2v_1 + u_1v_2) + du_2v_2.$$

Это значит, что для нашего самосопряженного линейного преобразования $\mathbf{A}$ и любых векторов-стрелок $\mathbf{u}$ и $\mathbf{v}$ верно, что $(\mathbf{A}(\mathbf{u}), \mathbf{v}) = (\mathbf{u}, \mathbf{A}(\mathbf{v}))$.

5.4. Собственные векторы самосопряженного линейного преобразования

Найдем собственные векторы самосопряженного преобразования, которое в некотором базисе имеет матрицу

$$\begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}.$$

Характеристическое уравнение

$$\begin{vmatrix} 5-\lambda & 2 \\ 2 & 2-\lambda \end{vmatrix} = (5-\lambda)(2-\lambda) - 4 = \lambda^2 - 7\lambda + 10 - 4 = \lambda^2 - 7\lambda + 6 = 0$$

имеет корни 1 и 6. Система уравнений для корня 1 состоит из двух пропорциональных уравнений

$$\begin{cases} 4x + 2y = 0 \\ 2x + y = 0, \end{cases}$$

в качестве решения которой выберем вектор-столбец $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

Для корня 6 имеем

$$\begin{cases} -x + 2y = 0 \\ 2x - 4y = 0 \end{cases}$$

с решением $\begin{pmatrix} 2 \\ -1 \end{pmatrix}$.

Случаен ли тот факт, что скалярное произведение этих собственных векторов равно нулю? Коротким рассуждением можно показать, что для самосопряженного линейного преобразования ортогональность собственных векторов, соответствующих неравным собственным значениям, имеет место всегда.

Теорема 5.2. *Пусть* λ *и* μ *собственные значения линейного преобразования* $\mathbf{A}$ *и* $\lambda \neq \mu$, *а* $\mathbf{u}$ *и* $\mathbf{v}$ *соответствующие этим собственным значениям собственные векторы.*
Тогда $(\mathbf{u}, \mathbf{v}) = 0$.

Запишем в виде диаграммы систему равенств.

$$\begin{array}{ccc} (\mathbf{A}(\mathbf{u}), \mathbf{v}) & = & (\mathbf{u}, \mathbf{A}(\mathbf{v})) \\ \| & & \| \\ (\lambda \mathbf{u}, \mathbf{v}) & = & (\mathbf{u}, \mu \mathbf{v}) \\ \| & & \| \\ \lambda(\mathbf{u}, \mathbf{v}) & = & \mu(\mathbf{u}, \mathbf{v}) \end{array}$$

Первая строка выражает самосопряженность преобразования, переход ко второй осуществляется благодаря тому, что $\mathbf{u}$ и $\mathbf{v}$ собственные векторы, а переход к третьей строке — по очевидному свойству скалярного произведения (см. упр. 5.1).

Равенство в третьей строке возможно, только если $\lambda = \mu$, либо если $(\mathbf{u}, \mathbf{v}) = \mathbf{0}$. Поскольку первое неверно, то верно второе, а значит, собственные векторы самосопряженного линейного преобразования, соответствующие различным собственным значениям, ортогональны.

Теорема доказана.

На этом хорошие свойства самосопряженных преобразований не кончаются.

Во-первых, все корни характеристического уравнения для симметричной матрицы — действительные числа, хотя для произвольной матрицы вполне возможны комплексные решения характеристического уравнения, а для них не могут быть найдены собственные векторы (если мы решим соответствующую систему линейных уравнений, то получим комплексные координаты векторов).

Во-вторых, оказывается, что собственных векторов самосопряженного преобразования всегда столько, какова размерность пространства, в котором преобразование действует.

Рассмотрим пример.

Пример 1. Найдем собственные векторы преобразования с матрицей

$$\begin{pmatrix} 0 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & 1 & -1 \\ \sqrt{2} & -1 & 1 \end{pmatrix}.$$

Решаем характеристическое уравнение:

$$\begin{vmatrix} -\lambda & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & 1-\lambda & -1 \\ \sqrt{2} & -1 & 1-\lambda \end{vmatrix} = (-\lambda)(1-\lambda)^2 - 2(1-\lambda) - 2(1-\lambda) + \lambda - 2 - 2 =$$

$$-\lambda^3+2\lambda^2+4\lambda-8 = -\lambda^2(\lambda-2)+4(\lambda-2) = (\lambda-2)(4-\lambda^2) = -(\lambda-2)^2(2+\lambda).$$

Это уравнение имеет два корня: $\lambda_1 = -2$ имеет кратность 1, $\lambda_2 = 2$ имеет кратность 2.

Первый корень дает нам систему

$$\begin{cases} 2x + \sqrt{2}y + \sqrt{2}z = 0 \\ \sqrt{2}x + 3y - z = 0 \\ \sqrt{2}x - y + 3z = 0, \end{cases}$$

которая имеет решение $\begin{pmatrix} -\sqrt{2} \\ 1 \\ 1 \end{pmatrix}$.

Второй же корень — кратный, т.е. в разложении характеристического многочлена на множители он содержится в двух скобках, можно сказать, что это два корня, которые случайно оказались равны. Соответствующая система уравнений содержит три пропорциональных уравнения

$$\begin{cases} -2x + \sqrt{2}y + \sqrt{2}z = 0 \\ \sqrt{2}x - y - z = 0 \\ \sqrt{2}x - y - z = 0, \end{cases}$$

а это, как мы помним, означает, что мы можем найти два линейно независимых (в данном случае это означает непропорциональных) решения системы, например $\mathbf{r}_1 = \begin{pmatrix} 1 \\ \sqrt{2} \\ 0 \end{pmatrix}$ и $\mathbf{r}_2 = \begin{pmatrix} 1 \\ 0 \\ \sqrt{2} \end{pmatrix}$.

Упражнение 5.3. Проверить, что всякая линейная комбинация решений $\mu_1\mathbf{r}_1 + \mu_2\mathbf{r}_2$ также будет решением системы, а значит, собственным вектором с собственным значением 2.

Выполнив упражнение, мы можем быть уверены, что решения системы представляют собой подпространство[1]. В данном случае его размерность равна двум, т.е. это плоскость. Базисом на этой плоскости будет пара $\mathbf{r}_1, \mathbf{r}_2$.

Как и на всякой другой, на этой плоскости мы можем выбрать ортогональный базис. В данном случае это сделать легко: векторы $\mathbf{r}_1 + \mathbf{r}_2 = \begin{pmatrix} 2 \\ \sqrt{2} \\ \sqrt{2} \end{pmatrix}$ и $\mathbf{r}_1 - \mathbf{r}_2 = \begin{pmatrix} 0 \\ \sqrt{2} \\ -\sqrt{2} \end{pmatrix}$ ортогональны[2]. Мы можем заметить теперь, что в нашем примере двум "случайно" равным корням "неслучайно" соответствуют два ортогональных решения системы, т.е. кратные корни "поставляют" ровно столько ортогональных собственных векторов самосопряженного преобразования, какова их кратность.

В заключение параграфа суммируем сказанное: у самосопряженного линейного преобразования всегда найдется ортонормированный базис, состоящий из собственных векторов. В следующей главе мы докажем это утверждение в общем случае.

Упражнение 5.4. Как из ортогонального базиса сделать ортонормированный?

---

[1] Определение подпространства будет дано в следующей главе.

[2] Есть также алгоритм ортогонализации любого базиса в пространстве любой размерности, но мы не будем его здесь приводить.

Глава 6

Линейные преобразования в евклидовых пространствах. Общий случай

6.1. Евклидовы пространства

Что можно было бы назвать прямым углом в четырехмерном пространстве? Мы, конечно, не можем применять там настоящие измерительные инструменты — транспортиры и угольники, но мы можем задать нечто похожее на транспортир. Этим нечто будет скалярное произведение.

Таким образом, наши определения будут следовать в порядке, противоположном порядку определений предыдущей главы. В трехмерном пространстве мы до всякой линейной алгебры знали прямые углы, и скалярное произведение попадало к нам в линейную алгебру извне, из области почти физически существующих векторов.

В пространстве большего числа измерений мы начинаем со скалярного произведения, поскольку никакого "извне" для абстрактного линейного пространства нет. Мы задаем в пространстве прямые углы, когда определяем в нем скалярное произведение: ортогональны те векторы, скалярное произведение которых равно нулю.

Скалярным произведением мы можем назвать любую функцию, обладающую теми свойствами, которые мы обнаружили у скалярного произведения векторов-стрелок.

Определение 6.1. *Скалярным произведением векторов в линейном пространстве* $\mathbf{V}$ *может быть любая функция* $\mathbf{S}(\mathbf{u},\mathbf{v})$, *удовлетворяющая следующим условиям:*

1) $\mathbf{S}(\mathbf{u},\mathbf{v}) = \mathbf{S}(\mathbf{v},\mathbf{u})$;
2) $\mathbf{S}(\mathbf{u}+\mathbf{u}',\mathbf{v}) = \mathbf{S}(\mathbf{u},\mathbf{v}) + \mathbf{S}(\mathbf{u}',\mathbf{v})$;
3) $\mathbf{S}(\lambda\mathbf{u},\mathbf{v}) = \lambda\mathbf{S}(\mathbf{u},\mathbf{v})$;
4) $\mathbf{S}(\mathbf{u},\mathbf{u}) > \mathbf{0}$.

Линейное пространство, в котором зафиксировано какое-то скалярное произведение, называется евклидовым. Это зафиксированное скалярное произведение обозначается обычно $(\mathbf{u},\mathbf{v})$.

Определение 6.2. *Базис* $\{\mathbf{e_1},\mathbf{e_2},\ldots,\mathbf{e_n}\}$ *в n-мерном евклидовом пространстве* $\mathbf{V}$ *называется ортонормированным, если* $(\mathbf{e_i},\mathbf{e_i}) = \mathbf{1}$ *и* $(\mathbf{e_i},\mathbf{e_j}) = \mathbf{0}$ *(при* $i \neq j$*).*

Теорема, параллельная теореме 1 предыдущей главы, оказывается существенно проще, поскольку свойство $(\mathbf{u}+\mathbf{u}',\mathbf{v}) = (\mathbf{u},\mathbf{v}) + (\mathbf{u}',\mathbf{v})$ введено в определение скалярного произведения, в то время как в предыдущей главе мы должны были его доказывать.

Теорема 6.1. *Пусть* $\mathbf{u}$ *и* $\mathbf{v}$ *векторы, а* $\begin{pmatrix} u_1 \\ \ldots \\ u_n \end{pmatrix}$ *и* $\begin{pmatrix} v_1 \\ \ldots \\ v_n \end{pmatrix}$ *их координатное выражение в некотором ортонормированном базисе* $\{\mathbf{e_1},\ldots,\mathbf{e_n}\}$. *Тогда* $(\mathbf{u},\mathbf{v}) = u_1v_1 + \ldots + u_nv_n$.

Доказательство. Из пунктов 1 и 2 определения скалярного произведения $(\mathbf{u},\mathbf{v}+\mathbf{v}') = (\mathbf{u},\mathbf{v}) + (\mathbf{u},\mathbf{v}')$ и $(\mathbf{u}+\mathbf{u}',\mathbf{v}) = (\mathbf{u},\mathbf{v}) + (\mathbf{u}',\mathbf{v})$.

Выразим векторы $\mathbf{u}$ и $\mathbf{v}$ через векторы базиса:

$\mathbf{u} = u_1\mathbf{e_1} + \ldots + u_n\mathbf{e_n}$;

$\mathbf{v} = v_1\mathbf{e_1} + \ldots + v_n\mathbf{e_n}$.

Перемножим теперь скалярно эти выражения:

$$\begin{aligned}
&(\mathbf{u},\mathbf{v}) = \\
&= (u_1v_1(\mathbf{e_1},\mathbf{e_1}) + u_1v_2(\mathbf{e_1},\mathbf{e_2}) + \ldots u_1v_n(\mathbf{e_1},\mathbf{e_n}) + \\
&+(u_2v_1(\mathbf{e_2},\mathbf{e_1}) + u_2v_2(\mathbf{e_2},\mathbf{e_2}) + \ldots u_2v_n(\mathbf{e_2},\mathbf{e_n}) + \\
&+ \ldots + \\
&+(u_nv_1(\mathbf{e_n},\mathbf{e_1}) + u_nv_2(\mathbf{e_n},\mathbf{e_2}) + \ldots u_nv_n(\mathbf{e_n},\mathbf{e_n}).
\end{aligned}$$

Поскольку базис $\{\mathbf{e_1}, \ldots, \mathbf{e_n}\}$ ортонормированный, все скалярные произведения на главной диагонали таблицы равны единице, а скалярные произведения вне главной диагонали таблицы равны нулю. Это и значит, что

$$(\mathbf{u}, \mathbf{v}) = u_1 v_1 + \ldots + u_n v_n,$$

причем мы доказали, что в такой форме выражается данное скалярное произведение в любом базисе, ортонормированном в смысле самого этого скалярного произведения.

➽ **Определение 6.3.** *Длиной вектора* $\mathbf{u}$ *в евклидовом пространстве* $\mathbf{V}$ *называется* $\sqrt{(\mathbf{u}, \mathbf{u})}$.

Длина вектора $\mathbf{u}$ обозначается $|\mathbf{u}|$. В любом ортонормированном базисе длина вычисляется по формуле $|\mathbf{u}| = \sqrt{u_1^2 + \ldots + u_n^2}$.

6.2. Замена ортонормированного базиса. Ортогональные матрицы

Пусть в евклидовом пространстве задан базис $\{\mathbf{e_1}, \ldots, \mathbf{e_n}\}$. Пусть $\{\mathbf{f_1}, \ldots, \mathbf{f_n}\}$ другой ортонормированный базис в том же пространстве с формулами перехода

$$\begin{array}{ccc} \mathbf{f_1} & \ldots & \mathbf{f_n} \\ \| & & \| \\ c_{11}\mathbf{e_1} & \ldots & c_{1n}\mathbf{e_1} \\ + & & + \\ \ldots & \ldots & \ldots \\ + & & + \\ c_{n1}\mathbf{e_n} & \ldots & c_{nn}\mathbf{e_n}. \end{array}$$

Будем обозначать буквой C матрицу перехода

$$\begin{pmatrix} c_{11} & \ldots & c_{1n} \\ \ldots & \ldots & \ldots \\ c_{n1} & \ldots & c_{nn} \end{pmatrix}.$$

Теорема 6.2. *Обратная к* C *матрица* C^{-1} *равна* C' *(транспонированной матрице* C*).*

Набросок *Доказательства.* Поскольку $\mathbf{f_i} = c_{1i}\mathbf{e_1} + \ldots + c_{ni}\mathbf{e_n}$ и векторы $\mathbf{f_i}$ имеют единичную длину и взаимно ортогональны, то, вычисляя

элементы произведения $U = C'C$, получим $u_{ii} = c_{1i}c_{1i} + \ldots + c_{ni}c_{ni} = (\mathbf{f_i}, \mathbf{f_i}) = 1$ и $u_{ij} = c_{1i}c_{1j} + \ldots + c_{ni}c_{nj} = (\mathbf{f_i}, \mathbf{f_i}) = 0$.

Это значит, что $U = C'C$ — единичная матрица. Для завершения доказательства надо показать, что CC' также единичная матрица. Строгое доказательство этого утверждения мы дадим в главе 8.

Определение 6.4. *Матрицы, обладающие свойством $CC' = C'C = E$, называются ортогональными.*

Матрица перехода от одного ортонормированного базиса к другому ортогональна.

6.3. Самосопряженные линейные преобразования

Определение 6.5. *Матрицы, обладающие свойством $a_{ij} = a_{ji}$, называются симметричными.*

В параграфе 4.5 главы 4 было показано, что если $\mathbf{A}$ линейное преобразование, которое в базисах $\{\mathbf{e_i}\}$ и $\{\mathbf{f_i}\}$ представляют матрицы A_e и A_f соответственно, а C матрица перехода от первого базиса ко второму, то матрицы преобразования связаны формулой $A_f = C^{-1}A_eC$.

В случае ортонормированных базисов $C^{-1} = C'$, поэтому связь матриц выражается также формулой $A_f = C'A_eC$. Далее в этой главе мы всегда будем говорить только об ортонормированных базисах.

Теорема 6.3. *Если матрица A_e симметрична, то и матрица A_f симметрична.*

Доказательство. Покажем сначала, что для любых матриц, которые можно перемножать, выполнено равенство

$$(MK)' = K'M',$$

т.е. транспонирование произведения матриц приводит к тому же результату, что и перемножение в обратном порядке транспонированных сомножителей. Выделим в сомножителях: i-ю строку матрицы M, j-й столбец матрицы K, j-ю строку матрицы K' и i-й столбец матрицы M' —

$$\begin{pmatrix} \ldots & \ldots & \ldots \\ m_{i1} & \ldots & m_{in} \\ \ldots & \ldots & \ldots \end{pmatrix} \begin{pmatrix} \ldots & k_{1j} & \ldots \\ \ldots & \ldots & \ldots \\ \ldots & k_{nj} & \ldots \end{pmatrix} \text{ и}$$

$$\begin{pmatrix} \dots & \dots & \dots \\ k_{1j} & \dots & k_{nj} \\ \dots & \dots & \dots \end{pmatrix} \begin{pmatrix} \dots & m_{i1} & \dots \\ \dots & \dots & \dots \\ \dots & m_{in} & \dots \end{pmatrix}.$$

Перемножение левой пары даст величину $m_{i1}k_{1j} + \dots + m_{in}k_{nj}$, которую следует поставить в матрицу-произведение на место ij, а в правой паре получим равную ей величину $k_{1j}m_{i1} + \dots + k_{nj}m_{in}$, которую следует поставить на место ji. Это и означает, что левое произведение даст матрицу, элементы которой равны расположенным симметрично элементам правого произведения, т.е.

$$(MK)' = K'M'.$$

Из только что доказанного следует[1], что для трех матриц

$$(MNK)' = K'N'M',$$

поскольку $(MNK)' = (NK)'M' = K'N'M'$.

Продолжим теперь доказательство теоремы. Нам надо доказать, что матрица $A_f = C'A_eC$ симметрична, т.е. равна своей транспонированной. Действительно, по только что доказанному, транспонированная матрица $(C'A_eC)'$ равна $C'A_e'(C')'$. Но A_e симметрична, поэтому $A_e = A_e'$, а равенство $(C')' = C$ выполняется вообще для всякой матрицы (двукратное отражение), поэтому $C'A_e'(C')' = C'A_eC$, т.е. транспонирование не изменило матрицу A_f, следовательно, она симметрична.

➽ **Определение 6.6.** *Линейное преобразование называется самосопряженным, если во всех ортонормированных базисах его матрица симметрична.*

По теореме 6.3 для этого достаточно, чтобы симметричной была его матрица хотя бы в одном ортонормированном базисе.

Дальнейшая наша цель — доказать, что собственные векторы самосопряженного преобразования образуют ортонормированный базис пространства. Для этого докажем несколько интересных утверждений.

Теорема 6.4.1. *Пусть* $\mathbf{A}$ *линейное преобразование. Если для любых двух векторов* $\mathbf{u}$ *и* $\mathbf{v}$ $(\mathbf{A}(\mathbf{u}), \mathbf{v}) = (\mathbf{u}, \mathbf{A}(\mathbf{v}))$, *то* $\mathbf{A}$ *самосопряженное преобразование.*

[1] Мы использовали здесь неявно равенство $(MN)K = M(NK)$. Подробнее об этом в главе 8.

Доказательство. Зафиксируем некоторый ортонормированный базис $\{\mathbf{e}_i\}$ и матрицу A в этом базисе.

Поскольку $(\mathbf{e}_i, \mathbf{A}(\mathbf{e}_j)) = a_{ij}$, $(\mathbf{A}(\mathbf{e}_i), \mathbf{e}_j) = a_{ji}$ и $(\mathbf{e}_i, \mathbf{A}(\mathbf{e}_j)) = (\mathbf{A}(\mathbf{e}_i), \mathbf{e}_j)$, то $a_{ij} = a_{ji}$. Таким образом, матрица A симметрична и $\mathbf{A}$ самосопряженное преобразование.

Теорема 6.4.2. *Если* $\mathbf{A}$ *самосопряженное преобразование, то для любых двух векторов* $\mathbf{u}$ *и* $\mathbf{v}$ *имеет место равенство* $(\mathbf{A}(\mathbf{u}), \mathbf{v}) = (\mathbf{u}, \mathbf{A}(\mathbf{v}))$.

Доказательство. Зафиксируем ортонормированный базис $\{\mathbf{e}_i\}$ и докажем утверждение для матрицы A, которую имеет преобразование в этом базисе, и векторов-столбцов u и v, соответствующих $\mathbf{u}$ и $\mathbf{v}$. Требуется доказать, что $(Au, v) = (u, Av)$. Левую часть равенства (Au, v) подсчитаем непосредственно:

$$\left(\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}, \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix}\right) =$$

$$= \left(\begin{pmatrix} a_{11}u_1 & + & a_{12}u_2 & + & \dots & + & a_{1n}u_n \\ a_{21}u_1 & + & a_{22}u_2 & + & \dots & + & a_{2n}u_n \\ \dots & & \dots & & \dots & & \dots \\ a_{n1}u_1 & + & a_{n2}u_2 & + & \dots & + & a_{nn}u_n \end{pmatrix}, \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix}\right) =$$

$$\begin{array}{lllllllll} = & a_{11}u_1v_1 & + & a_{12}u_2v_1 & + & \dots & + & a_{1n}u_nv_1 & + \\ + & a_{21}u_1v_2 & + & a_{22}u_2v_2 & + & \dots & + & a_{2n}u_nv_2 & + \\ & \dots & & \dots & & \dots & & \dots & \\ + & a_{n1}u_1v_n & + & a_{n2}u_2v_n & + & \dots & + & a_{nn}u_nv_n. & \end{array} \tag{6.1}$$

При подсчете правой части (u, Av) мы несколько схитрим: поскольку $(u, Av) = (Av, u)$ (пункт 1 определения скалярного произведения), то результат расчета (Av, u) можно получить из предыдущего расчета (Au, v), подставив в итоговую сумму (6.1) одновременно вместо всех букв u буквы v, а вместо букв v буквы u:[2]

$$(u, Av) =$$

[2] При подготовке текста автор именно так и сделал, воспользовавшись функцией “Replace” текстового редактора.

$$\begin{array}{cccccccccc} = & a_{11}v_1u_1 & + & a_{12}v_2u_1 & + & \dots & + & a_{1n}v_nu_1 & + \\ + & a_{21}v_1u_2 & + & a_{22}v_2u_2 & + & \dots & + & a_{2n}v_nu_2 & + \\ & \dots & & \dots & & \dots & & \dots & \\ + & a_{n1}v_1u_n & + & a_{n2}v_2u_n & + & \dots & + & a_{nn}v_nu_n. & \end{array} \qquad (6.2)$$

Сравнение выражений (6.1) и (6.2) показывает, что они равны, если матрица симметрична и $a_{ij} = a_{ji}$, что имеет место по условию теоремы.

Теорема доказана.

6.4. Собственные векторы самосопряженного линейного преобразования

Мы можем повторить приведенное в предыдущей главе общее доказательство того факта, что собственные векторы самосопряженного линейного преобразования, соответствующие различным собственным значениям, ортогональны (теорема 5.2, параграф 5.4 главы 5), однако это не поможет нам достичь нашей цели — доказать, что самосопряженное преобразование имеет базис из собственных векторов. Действительно, если характеристическое уравнение имеет кратные корни, то последнего утверждения недостаточно.

Чтобы справиться с этой трудностью, мы введем два новых понятия.

➾ **Определение 6.7.** *Назовем линейное пространство* $\mathbf{W}$ *подпространством линейного пространства* $\mathbf{V}$, *если все элементы* $\mathbf{W}$ *являются элементами* $\mathbf{V}$ *(в математической традиции принято считать, что подпространство может и совпадать со всем пространством; если же оно "меньше", т.е. имеются элементы* $\mathbf{V}$, *не принадлежащие* $\mathbf{W}$, *то такое подпространство называется собственным).*

➾ **Определение 6.8.** *Назовем подпространство* $\mathbf{W}$ *инвариантным для преобразования* $\mathbf{A}$, *если при этом преобразовании образ любого вектора из* $\mathbf{W}$ *также принадлежит* $\mathbf{W}$. .

Пусть $\mathbf{u}$ вектор из $\mathbf{V}$. Рассмотрим множество всех векторов, ему ортогональных. Обозначим это множество через $\mathbf{V_u}$.

Теорема 6.5. $\mathbf{V_u}$ *представляет собой линейное пространство.*

Доказательство. Надо проверить, что выполнены все пункты определения линейного пространства, которое было дано в начале четвертой главы.

1. Если $\mathbf{w}$ и $\mathbf{v}$ векторы пространства $\mathbf{V_u}$, то и $\mathbf{w}+\mathbf{v}$ также элемент $\mathbf{V_u}$.

Это действительно так: поскольку $\mathbf{w}$ и $\mathbf{v}$ ортогональны $\mathbf{u}$, то $(\mathbf{w},\mathbf{u}) = (\mathbf{v},\mathbf{u}) = \mathbf{0}$. Но тогда $(\mathbf{w}+\mathbf{v},\mathbf{u}) = (\mathbf{w},\mathbf{u}) + (\mathbf{v},\mathbf{u})$ также равно нулю. Это значит, что сумма $\mathbf{w}+\mathbf{v}$ ортогональна вектору $\mathbf{u}$, т.е.принадлежит $\mathbf{V_u}$.

2. Если $\mathbf{v}$ принадлежит $\mathbf{V_u}$, а λ действительное число, то и $\lambda\mathbf{v}$ также принадлежит $\mathbf{V_u}$, поскольку $(\lambda\mathbf{v},\mathbf{u}) = \lambda(\mathbf{v},\mathbf{u}) = 0$.

3. Нулевой вектор $\mathbf{0}$ принадлежит $\mathbf{V_u}$, поскольку $(\mathbf{0},\mathbf{u}) = \mathbf{0}$.

4. Если $\mathbf{v}$ принадлежит $\mathbf{V_u}$, то и $-\mathbf{v}$ принадлежит $\mathbf{V_u}$, поскольку $(-\mathbf{v},\mathbf{u}) = -(\mathbf{v},\mathbf{u}) = \mathbf{0}$.

Теорема доказана

Теорема 6.6. *Если* $\mathbf{u}$ *собственный вектор самосопряженного преобразования* $\mathbf{A}$, *то подпространство* $\mathbf{V_u}$ *инвариантно для* $\mathbf{A}$.

Доказательство. Если $\mathbf{v}$ принадлежит $\mathbf{V_u}$, то $(\mathbf{v},\mathbf{u}) = 0$. Но тогда и $(\mathbf{A}(\mathbf{v}),\mathbf{u}) = 0$, поскольку $(\mathbf{A}(\mathbf{v}),\mathbf{u}) = (\mathbf{v},\mathbf{A}(\mathbf{u})) = (\mathbf{v},\lambda_u\mathbf{u}) = \lambda_u(\mathbf{v},\mathbf{u}) = 0$, следовательно, $\mathbf{A}(\mathbf{v})$ принадлежит $\mathbf{V_u}$, что по определению означает, что $\mathbf{V_u}$ инвариантно для $\mathbf{A}$.

Теорема доказана.

Теорема 6.7. *Все собственные значения самосопряженного преобразования действительные числа.*

Доказательство этого факта мы не будем приводить, поскольку оно использует комплексное линейное пространство, а разработка этого понятия потребовала бы от нас неоправданно больших усилий.

Мы имеем теперь все, что требуется для доказательства наличия базиса из собственных векторов самосопряженного преобразования.

Доказательство проводится по индукции сведением к более низкой размерности пространства.

Теорема 6.8. *Пусть* $\mathbf{A}$ *самосопряженное линейное преобразование пространства* $\mathbf{V}$. *Тогда в* $\mathbf{V}$ *имеется базис из собственных векторов пространства* $\mathbf{V}$.

Доказательство. Поскольку все корни характеристического уравнения для $\mathbf{A}$ действительны, то даже в случае, если все они совпадают, т.е. кратность корня равна размерности пространства $\mathbf{V}$, мы обязательно найдем хотя бы один собственный вектор $\mathbf{u}$.

Рассмотрим его ортогональное дополнение $\mathbf{V_u}$. Забудем на время про объемлющее пространство $\mathbf{V}$ и посмотрим, как $\mathbf{A}$ действует на векторы $\mathbf{V_u}$. Поскольку пространство $\mathbf{V_u}$ инвариантно, то $\mathbf{A}$ можно рассматривать как преобразование $\mathbf{V_u}$. Поскольку $(\mathbf{A}(\mathbf{w}), \mathbf{v}) = (\mathbf{w}, \mathbf{A}(\mathbf{v}))$ выполняется и для векторов $\mathbf{V_u}$, то преобразование $\mathbf{A}$ остается самосопряженным в пространстве $\mathbf{V_u}$ (теорема 6.4.1).

Мы теперь можем решить характеристическое уравнение для нового преобразования, найти его собственные значения, хотя бы один собственный вектор и рассмотреть ортогональное дополнение уже к этому новому вектору.

Последовательное проведение этой процедуры приведет в конце концов к одномерному подпространству с последним собственным вектором в качестве его базиса. Все полученные последовательно векторы будут собственными для самого преобразования $\mathbf{A}$, их количество будет равно n и все они будут взаимно ортогональны.

Для того чтобы считать доказательство теоремы законченным, математик потребует только доказать, что любые n взаимно ортогональных векторов в n-мерном пространстве линейно независимы, а значит, образуют базис.

Пусть $\mathbf{v_1}, \ldots, \mathbf{v_n}$ взаимно ортогональные векторы.

Предположим, что они линейно зависимы, т.е. существует их ненулевая линейная комбинация, равная нулевому вектору:

$$\lambda_1 \mathbf{v_1} +, \ldots, \lambda_n \mathbf{v_n} = \mathbf{0},$$

причем некоторое $\lambda_i \neq 0$. Умножим последнее равенство скалярно на $\mathbf{v_i}$. Получим

$$\lambda_1(\mathbf{v_1}, \mathbf{v_i}) +, \ldots, +\lambda_i(\mathbf{v_i}, \mathbf{v_i}) + \ldots + \lambda_n(\mathbf{v_n}, \mathbf{v_i}) = 0.$$

В этой сумме все скалярные произведения, кроме одного, равны нулю, поскольку векторы ортогональны. Это значит, что

$$\lambda_i(\mathbf{v_i}, \mathbf{v_i}) = 0,$$

что невозможно, поскольку мы выбрали $\lambda_i \neq 0$, а $(\mathbf{v_i}, \mathbf{v_i}) = \mathbf{1}$.

Тем самым **теорема доказана**.

Глава 7

Линейная алгебра в факторном анализе

7.1. Метод главных компонент

В предыдущих главах мы изучали самосопряженные преобразования в евклидовых пространствах. Мы обнаружили, что у каждого такого преобразования можно найти базис из собственных векторов, т.е. векторов, которые преобразуются самым простым образом — удлиняются или укорачиваются, но не поворачиваются.

Мы обнаружили также, что в каждом базисе линейное преобразование может быть задано матрицей и что при переходе от одного базиса к другому матрицы преобразуются по формуле $A_f = C'A_eC$.

Если $\mathbf{f}_i$ базис из собственных векторов, а C матрица перехода от какого-то базиса $\mathbf{e}_i$ к базису $\mathbf{f}_i$, то в базисе $\mathbf{f}_i$ матрица преобразования будет диагональной.

Матрицами могут быть заданы и другие интересные объекты. Мы рассмотрим ниже два вида матриц — матрицы выборочных ковариаций и корреляций. Не давая пока определений, заметим, что выборочные ковариации и корреляции столь же осмысленный объект, заданный в n-мерном линейном пространстве, что и линейное преобразование. Они описывают некоторые важные характеристики выборок — совокупностей эмпирических результатов разнообразной природы.

Интересующие нас матрицы, во-первых, симметричны, а во-вторых, так же как и матрицы линейного преобразования, меняются при замене базиса. При переходе от одного ортонормированного базиса к друго-

му ортонормированному новая и старая матрицы связаны формулой $K_f = C'K_eC$, т.е. той же самой формулой, что и матрица линейного преобразования.

Это позволяет проделать с матрицами выборочных ковариаций и корреляций K те же самые операции, что и в случае линейного преобразования. Мы можем понимать эти операции так: рассмотрим самосопряженное линейное преобразование, которое в данном базисе имеет ту самую матрицу — например, рассчитанную в данном базисе матрицу выборочных ковариаций K. У выборочных ковариаций нет собственных векторов, поскольку ковариации ничего не преобразуют, но у имеющего ту же матрицу преобразования можно найти базис из собственных векторов.

Однако поскольку формулы пересчета матриц при замене базиса совпадают у линейных преобразований и выборочных ковариаций, то матрица выборочных ковариаций в новом базисе также будет совпадать с матрицей преобразования и, следовательно, будет диагональной. Это ее качество имеет вполне серьезную интерпретацию с точки зрения выборочных ковариаций. Более того, в терминах выборочных ковариаций интерпретируются и собственные значения, которые соответствуют векторам собственного базиса линейного преобразования.

Пример 1. В нашем примере количество испытуемых и размерность пространства выбраны минимальными, чтобы расчеты не были слишком сложными. В практических приложениях размерности матриц значительно больше.

Пусть четверо студентов получили по результатам тестирования по две оценки, первая характеризует их успешность по математике, вторая — по психологии:

$$\begin{pmatrix} 1 \\ \sqrt{2} \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ -\sqrt{2} \end{pmatrix}.$$

Как мы говорили, такая совокупность результатов называется выборкой.

Расчитанная по данной выборке матрица выборочных ковариаций имеет такой вид:

$$\begin{pmatrix} \dfrac{1^2+1^2+(-1)^2+(-1)^2}{4} & \dfrac{1\cdot\sqrt{2}+1\cdot 0+(-1)\cdot 0+(-1)\cdot(-\sqrt{2})}{4} \\ \dfrac{1\cdot\sqrt{2}+1\cdot 0+(-1)\cdot 0+(-1)\cdot(-\sqrt{2})}{4} & \dfrac{(\sqrt{2}^2+0^2+0^2+(-\sqrt{2})^2}{4} \end{pmatrix}.$$

В левом верхнем углу матрицы стоит сумма квадратов оценок по первой дисциплине деленная на их количество, в правом нижнем —

сумма квадратов оценок по второй дисциплине, также деленная на четыре. В левом верхнем и правом нижнем углах стоят одинаковые суммы попарных произведений оценок по первой и второй дисциплинам, также деленные на четыре. Произведя вычисления, получаем матрицу

$$\begin{pmatrix} 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix}.$$

Это матрица выборочных ковариаций в базисе

$$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix};\ e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

соответствующем исходным оценкам по дисциплинам.

Решая характеристическое уравнение

$$\begin{vmatrix} 1-\lambda & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 1/2 = 0,$$

получаем два корня:

$$\lambda_1 = 1 + \frac{1}{\sqrt{2}} \text{ и } \lambda_2 = 1 - \frac{1}{\sqrt{2}},$$

которым соответствуют собственные векторы имеющего ту же матрицу линейного преобразования

$$f_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \text{ и } f_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Эти выражения задают матрицу перехода от старого базиса к новому

$$C = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Мы можем рассматривать пары оценок четырех испытуемых как четыре вектора на плоскости, имеющих в старом базисе координаты

$$a_e = \begin{pmatrix} 1 \\ \sqrt{2} \end{pmatrix}, b_e = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, c_e = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, d_e = \begin{pmatrix} -1 \\ -\sqrt{2} \end{pmatrix}.$$

Для каждого из них мы можем получить координаты в базисе $\{\mathbf{f_i}\}$ по формулам перехода с матрицей $C^{-1} = C'$:

$$a_f = C'a_e,\ b_f = C'b_e,\ c_f = C'c_e,\ d_f = C'd_e.$$

Расчитываем новые координаты:

$$a_f = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 \\ \sqrt{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} + 1 \\ \frac{1}{\sqrt{2}} - 1 \end{pmatrix}.$$

Аналогично получаем

$$b_f = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}; c_f = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \text{ и } d_f = \begin{pmatrix} -\frac{1}{\sqrt{2}} - 1 \\ -\frac{1}{\sqrt{2}} + 1 \end{pmatrix}.$$

Если теперь посчитать матрицу выборочных ковариаций в новом базисе, то она окажется диагональной :

$$k_{11} = \tfrac{1}{4}((\tfrac{1}{\sqrt{2}} + 1)^2 + (\tfrac{1}{\sqrt{2}})^2 + (-\tfrac{1}{\sqrt{2}})^2 + (-\tfrac{1}{\sqrt{2}} - 1)^2) =$$
$$= \tfrac{1}{4}(\tfrac{1}{2} + \sqrt{2} + 1 + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \sqrt{2} + 1) = 1 + \tfrac{1}{\sqrt{2}},$$

$$k_{12} = k_{21} =$$
$$= \tfrac{1}{4}((\tfrac{1}{\sqrt{2}}+1)(\tfrac{1}{\sqrt{2}}-1) + \tfrac{1}{\sqrt{2}}\tfrac{1}{\sqrt{2}} + (-\tfrac{1}{\sqrt{2}})(-\tfrac{1}{\sqrt{2}}) + (-\tfrac{1}{\sqrt{2}}-1)(-\tfrac{1}{\sqrt{2}}+1)) =$$
$$= \tfrac{1}{4}(\tfrac{1}{2} - 1 + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} - 1) = 0,$$

$$k_{22} = \tfrac{1}{4}((-\tfrac{1}{\sqrt{2}} + 1)^2 + (\tfrac{1}{\sqrt{2}})^2 + (-\tfrac{1}{\sqrt{2}})^2 + (\tfrac{1}{\sqrt{2}} - 1)^2) =$$
$$= \tfrac{1}{4}(\tfrac{1}{2} - \sqrt{2} + 1 + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} - \sqrt{2} + 1) = 1 - \tfrac{1}{\sqrt{2}}.$$

Матрица, следовательно, имеет вид

$$K_f = \begin{pmatrix} 1 + \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 - \frac{1}{\sqrt{2}} \end{pmatrix}.$$

с собственными значениями на главной диагонали.

Элементы матрицы, лежащие вне главной диагонали, характеризуют связь соответствующих величин. Если в исходном базисе наша матрица имела в соответствующих местах $\frac{1}{\sqrt{2}}$, то это означает довольно сильную связь между оценками по математике и психологии, которую

продемонстрировали протестированные студенты. Подробно о корреляциях мы будем говорить в части 4 нашей книги. Нулевое значение элементов вне главной диагонали в матрице, соответсвующей новому базису, означает полное отсутствие связи (независимость) факторов $\mathbf{f}_1$ и $\mathbf{f}_2$.

После выявления независимых факторов наступает важный момент их интерпретации. В нашем упрощенном примере сделать это довольно просто: фактор $\mathbf{f}_1$ — это общие способности. Чем большую величину имеет первая координата испытуемого, тем более высокую оценку он имеет по обоим предметам. Второй фактор отвечает за склонности студента к техническим или гуманитарным дисциплинам: чем больше соответствующая координата, тем больше ожидаемая оценка по математике и меньше по психологии.

На рис. 7.1 изображена диаграмма рассеяния наших "экспериментальных" данных. Высокая корреляционная связь исходных переменных на рисунке обнаруживается в некоторой вытянутости облака точек в направлении диагонали между базисными векторами $\mathbf{e}_1$ и $\mathbf{e}_2$. Отсутствие корреляционной связи между факторами проявляется в отсутствии подобной "диагональной" тенденции по отношению к новым осям.

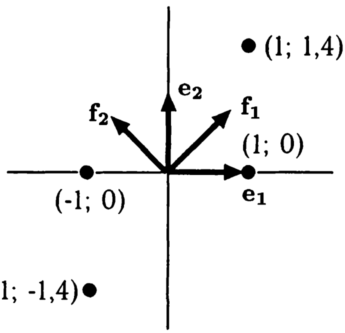


Рис. 7.1. Диаграмма рассеяния "экспериментальной" выборки

Явно неравная степень вытянутости вдоль новых осей напрямую связана с величиной соответствующих собственных значений. У первого фактора оно приблизительно равно 1,7, у второго 0,3 — "облако" на диаграмме рассеяния заметно вытянуто вдоль оси $\mathbf{f}_1$.

Внимательный читатель, возможно, заметил, что выборки нашего примера 1 обладали явно неслучайными качествами. Во-первых, сумма всех четырех векторов равна нулю. Во-вторых, на главной диагонали матрицы стоят единицы, а это значит, что средние квадраты переменных (сумма квадратов, деленная на количество испытуемых) равны единице.

Что касается первого свойства, то выборки, им обладающие, называются центрированными. Ковариации и корреляции расчитываются только для таких выборок.

Выборки, обладающие вторым свойством, называются стандартизованными. Зачем требуется стандартизация выборки, видно из следующего примера.

Пример 2. Пусть дана выборка, состоящая из данных тестирования четырёх испытуемых, где первая компонента измеряет количество успешно написанных контрольных работ по математике, а вторая - количество часов прослушанных лекций по психологическим дисциплинам за все время обучения:

$$\begin{pmatrix} 5 \\ 228 \end{pmatrix}, \begin{pmatrix} 5 \\ 214 \end{pmatrix}, \begin{pmatrix} 3 \\ 214 \end{pmatrix}, \begin{pmatrix} 3 \\ 200 \end{pmatrix}.$$

Как мы упоминали, для того чтобы считать ковариацию, выборка должна быть центрирована. Среднее значение первой координаты по четырем испытуемым равно 4, среднее значение по второй координате равно 214. Вычитаем эти значения. После центрирования выборка приобретает следующий вид:

$$\begin{pmatrix} 1 \\ 14 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ -14 \end{pmatrix}.$$

Матрица ковариаций, вычисленная по выборке, имеет следующий вид:

$$\begin{pmatrix} 1 & 7 \\ 7 & 98 \end{pmatrix}.$$

Характеристическое уравнение $(98-\lambda)(1-\lambda)-49 = \lambda^2-99\lambda+49 = 0$ имеет корни 0,5 и 98,5, с собственными векторами $\begin{pmatrix} 1 \\ 14{,}1 \end{pmatrix}$ и $\begin{pmatrix} -14{,}1 \\ 1 \end{pmatrix}$.

Независимые факторы практически совпадают с исходными переменными.

Можно задать вопрос, что будет, если продолжительность прослушанных лекций измерять в секундах или, наоборот, в годах. Оказывается, результат применения нашего метода существенно изменится. Посмотрим, что будет, если 14 часов выразить в долях семестра. Будем считать, что в среднем за семестр лекционный курс продолжается 35 часов, т.е. 14 часов это 0,4 семестра

$$\begin{pmatrix} 1 \\ 0{,}4 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ -0{,}4 \end{pmatrix}.$$

Матрица ковариаций, вычисленная по выборке, имеет следующий вид:

$$\begin{pmatrix} 1 & 0{,}2 \\ 0{,}2 & 0{,}08 \end{pmatrix}.$$

Характеристическое уравнение $(0{,}08-\lambda)(1-\lambda)-0{,}04=\lambda^2-1{,}08\lambda+0{,}04=0$ имеет корни, приблизительно равные 0,4 и 1,04, им соответствуют собственные векторы (с некоторой погрешностью округления) $\begin{pmatrix}5\\1\end{pmatrix}$ и $\begin{pmatrix}1\\-5\end{pmatrix}$.

Изменение единиц измерения вызвало существенное изменение результатов факторного анализа. К сожалению, неопределенность такого рода почти всегда сопутствует психологическим исследованиям, в отличие, скажем, от физических, где единицы измерения всегда согласованы.

В каких сопоставимых единицах можно было бы сравнивать, скажем, среднюю продолжительность безотрывного письма и количество пропусков гласных букв в письменных изложениях младших школьников?

Во всех случаях, когда подобные проблемы не имеют какого-то содержательно ясного решения, проводится стандартизация данных. Числа, которые помещались на главной диагонали матрицы выборочных ковариаций — выборочные дисперсии — служат нормирующими показателями.

➽ **Определение 7.1.** *Выборочной дисперсией центрированной переменной называется сумма квадратов ее значений, деленная на количество этих значений, т.е. на количество элементов выборки.*

В нашем примере для количества часов по психологии выборочная дисперсия составляет $(14^2+0^2+0^2+(-14)^2)/4=98$.

Для той же величины, измеренной в долях семестра, выборочная дисперсия составляет $(0{,}4^2+0^2+0^2+(-0{,}4)^2)/4=0{,}08$.

➽ **Определение 7.2.** *Стандартизовать переменную по данной выборке означает расчитать выборочную дисперсию данной переменной и каждое значение переменной поделить на квадратный корень из этой выборочной дисперсии.*

В первом случае $\sqrt{98}\approx 9{,}9$. Приблизительно 14/9,9 равно 1,414, поэтому стандартизация переменной преобразует значения 14, 0, 0, 14 в значения 1,414, 0, 0, -1,414.

Для той же величины, измеренной в долях семестра, выборочная дисперсия равна $\sqrt{0{,}08}\approx 0{,}28$. а $0{,}04/0{,}28=1{,}428$, и выборка приобретает похожий вид: 1,428, 0, 0, -1,428. Можно убедиться, что расхождения вызваны только ошибками округления.

Действительно, пусть $\{a;\ b;\ c;\ d\}$ значения переменной в какой-то выборке размера 4, а $\{\lambda a;\ \lambda b;\ \lambda c;\ \lambda d\}$ значения той же величины, измеренной в других единицах. Выборочная дисперсия

$$D = (a^2 + b^2 + c^2 + d^2)/4.$$

Стандартизованная по выборке переменная в первом случае имеет вид

$$\{a/\sqrt{D};\ b/\sqrt{D};\ c/\sqrt{D};\ d/\sqrt{D}\}.$$

Для второй переменной $\{\lambda a;\ \lambda b;\ \lambda c;\ \lambda d\}$, имеем

$$D' = (\lambda^2 a^2 + \lambda^2 b^2 + \lambda^2 c^2 + \lambda^2 d^2)/4 = \lambda^2 D.$$

При делении

$$\{\lambda a/\sqrt{\lambda^2 D};\ \lambda b/\sqrt{\lambda^2 D};\ \lambda c/\sqrt{\lambda^2 D};\ \lambda d/\sqrt{\lambda^2 D}\}$$

λ выносится из-под знака корня в знаменателе и сокращается с числителем, в результате чего получаем в точности равную выборку

$$\{a/\sqrt{D};\ b/\sqrt{D};\ c/\sqrt{D};\ d/\sqrt{D}\}.$$

Стандартизация — это только смена единиц измерения. Если нет содержательных соображений, диктующих выбор единиц, то разумнее всего избрать единицы измерения, связанные с разбросом результатов по данной переменной. Наиболее удобным в математическом смысле масштабом служит корень квадратный из дисперсии.

Определение 7.3. *Матрица ковариаций, расчитанная по стандартизованным переменным, называется матрицей корреляций.*

Подробнее о дисперсии, ковариации и корреляции мы будем говорить в четвертой части книги.

Наша выборка после стандартизации приобретает следующий вид:

$$\begin{pmatrix} 1 \\ \sqrt{2} \end{pmatrix},\ \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ \begin{pmatrix} -1 \\ 0 \end{pmatrix},\ \begin{pmatrix} -1 \\ -\sqrt{2} \end{pmatrix}.$$

Точно с теми же числами мы имели дело в примере 1 (отметим, что вообще случай двух переменных не может дать существенно иной картины). Соответствующая нашей выборке матрица корреляций такова:

$$\begin{pmatrix} 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix},$$

а ее собственные векторы

$$f_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \text{ и } f_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Первый фактор может быть интерпретирован как общее усердие студента, второй связан с наличием специальных математических/гуманитарных способностей (этот вымышленный упрощенный пример не следует принимать за научный анализ реальной ситуации).

7.2. Суммарная дисперсия. Доля фактора в суммарной дисперсии

Выше было дано определение выборочной дисперсии центрированной переменной. Эта дисперсия характеризует разброс значений данной переменной.

Общая дисперсия выборки складывается из выборочных дисперсий переменных и характеризует разброс точек на плоскости, заданных значениями переменных. Рассмотрим выборку примера 1:

$$\begin{pmatrix} 1 \\ \sqrt{2} \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ -\sqrt{2} \end{pmatrix}.$$

Общая дисперсия равна сумме выборочных дисперсий первой и второй переменных. Если переменные были стандартизованы, то общая дисперсия равна сумме единиц (единице равна дисперсия стандартизованной переменной), а число этих единиц равно количеству перменных.

В результате факторного анализа была получена диагональная матрица

$$\begin{pmatrix} 1 + \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 - \frac{1}{\sqrt{2}} \end{pmatrix},$$

у которой на главной диагонали стоят выборочные дисперсии новых переменных, а именно пересчитанных координат в новом базисе из собственных векторов $\{\mathbf{f_1}, \mathbf{f_2}\}$. Подробнее: новое представление выборки в

базисе факторов

$$\begin{pmatrix} \frac{1}{\sqrt{2}}+1 \\ \frac{1}{\sqrt{2}}-1 \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}}-1 \\ -\frac{1}{\sqrt{2}}+1 \end{pmatrix}.$$

дает выборочные дисперсии переменных $1+\frac{1}{\sqrt{2}}$ для первого фактора и $1-\frac{1}{\sqrt{2}}$ для второго фактора. Как видим, суммарная дисперсия выборки по-прежнему равна двум.

Упражнение 7.1. Доказать, что суммарная дисперсия выборки равна сумме квадратов длин входящих в выборку векторов.

Если читатель справился с упражнением, то сохранение суммарной дисперсии при замене базиса перестало его удивлять, поскольку сумма квадратов длин зависит только от векторов, а не от базисов.

По доле общей дисперсии, приходящейся на (выборочную) дисперсию переменной, соответствующей фактору, можно судить о важности данного фактора. В нашем примере 1 на фактор общих способностей приходится

$$\frac{1+\frac{1}{\sqrt{2}}}{2},$$

т.е. 85% общей дисперсии, а на фактор специальных способностей — лишь оставшиеся 15% общей дисперсии.

Пример 3. На рисунке 7.2 показана диаграмма рассеяния смоделированной с помощью датчика случайных чисел двумерной выборки из 100 испытуемых. Корреляция между исходными переменными равна 0,85. Чем больше значение переменной x_1, тем в среднем больше значение x_2. Связь переменных очень сильная. Факторный анализ в случае двух переменных всегда дает собственный базис, повернутый по отношению к исходному на 45 градусов (уже в случае трех переменных конфигурации факторов могут быть очень разнообразны).

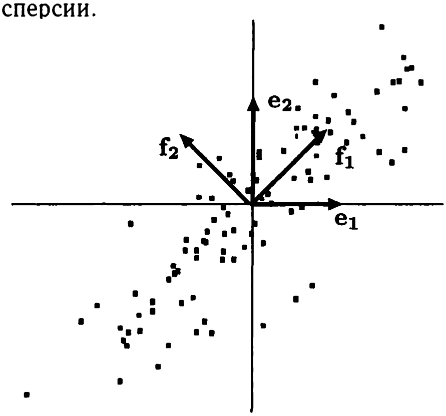


Рис. 7.2. Диаграмма рассеяния. Коэффициент корреляции равен 0,85

Собственные значения матрицы корелляций равны 1,85 и 0,15. "Облако" точек сильно вытянуто вдоль биссектрисы первого квадранта. Доли дисперсии, приходящиеся на факторы, равны, соответственно, 0,925 и 0,075.

Пример 4. На рисунке 7.3 показана аналогичная диаграмма рассеяния, но с корреляцией между исходными переменными, равной –0,5. В этом случае, чем больше значение переменной x_1, тем в среднем меньше значение x_2, и наоборот, чем, меньше значение переменной x_1, тем в среднем больше значение x_2. Связь переменных не так сильна, как в предыдущем примере. Собственные значения матрицы корелляций равны 1,5 и 0,5. "Облако" точек вытянуто вдоль биссектрисы второго квадранта несколько слабее,чем в предыдущем примере. Доли дисперсии, приходящиеся на факторы, равны, соответственно, 0,75 и 0,25.

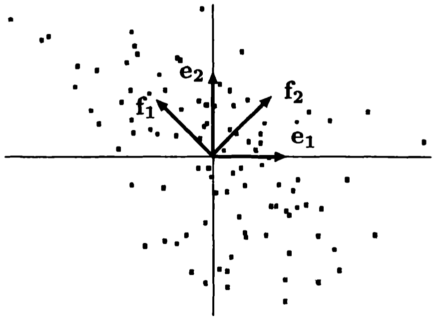


Рис. 7.3. Диаграмма рассеяния. Коэффициент корреляции равен $-0,5$

Пример 5. На рисунке 7.4 показана диаграмма рассеяния выборки с корреляцией между исходными переменными, равной 0,25. В этом случае, в принципе, чем больше значение переменной x_1, тем в среднем больше значение x_2, но связь переменных довольно слаба, о чем свидетельствует слаба вытянутость "облака" точек. Собственные значения матрицы корелляций равны 1,25 и 0,75. Доли дисперсии, приходящиеся на факторы, равны, соответственно, 0,625 и 0,375.

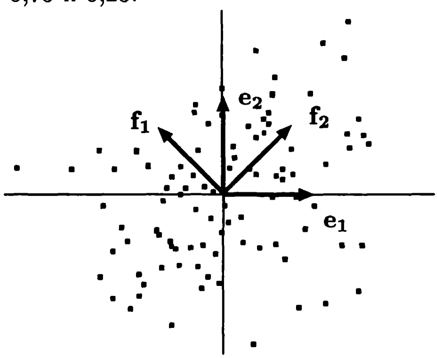


Рис. 7.4. Диаграмма рассеяния. Коэффициент корреляции равен 0,25

Глава 8

Метод главных компонент в общем случае

8.1. Элементы алгебры матриц

В главе 4 мы определили единственную операцию над матрицами — их умножение. Теперь мы дадим полную систему необходимых определений и докажем некоторые новые утверждения.

Определение умножения матриц мы не будем повторять (см. главы 3 и 4), поскольку после многократного использования читатель, наверное, его уже усвоил и запомнил.

Определение 8.1. *Суммой двух матриц*

$$\begin{pmatrix} a_{11} & \dots & a_{1k} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nk} \end{pmatrix} \text{ и } \begin{pmatrix} b_{11} & \dots & b_{1k} \\ \dots & \dots & \dots \\ b_{n1} & \dots & b_{nk} \end{pmatrix}$$

называется матрица

$$\begin{pmatrix} a_{11}+b_{11} & \dots & a_{1k}+b_{1k} \\ \dots & \dots & \dots \\ a_{n1}+b_{n1} & \dots & a_{nk}+b_{1k} \end{pmatrix}.$$

Определение 8.2. *Произведение действительного числа λ и матрицы A определяется так:*

$$\lambda \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nk} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \dots & \lambda a_{1k} \\ \dots & \dots & \dots \\ \lambda a_{n1} & \dots & \lambda a_{nk} \end{pmatrix}.$$

Операции над матрицами обладают рядом легко проверяемых свойств:

1) $\lambda(A + B) = \lambda A + \lambda B$;

2) $(\lambda\mu)(A) = \lambda(\mu A)$;

3) $(A + B)C = AC + BC; A(B + C) = AB + AC$;

и не столь очевидным свойством

4) $A(BC) = (AB)C$.

Это свойство мы использовали в главе 6. Так же как и предыдущие, оно проверяется непосредственно, но требует довольно громоздких выкладок, которые мы здесь приводить не будем.

Складывать можно только матрицы одинакового размера, для перемножения требуется равенство горизонтального размера первого сомножителя и вертикального размера второго. Умножение матрицы на вектор также является частным случаем умножения матриц, просто размер второго сомножителя $1 \times n$.

➽ **Определение 8.3.** *(Повторение определения 6 главы 4.) Единичной матрицей порядка n называется матрица $n \times n$ с единицами на главной диагонали и нулями в прочих местах. Единичная матрица обозначается буквой E, ее порядок обычно задается контекстом употребления.*

5) Для любой матрицы A произведения AE и EA равны исходной матрице E. (Если A не квадратная матрица, то это также верно, но при умножении справа и слева надо использовать матрицы E разного порядка.)

➽ **Определение 8.4.** *(Повторение определения 8 главы 4.) Обратной к матрице A называется матрица B, такая, что $AB = BA = E$. Обратная матрица обычно обозначается A^{-1}.*

Этих алгебраических свойств матриц достаточно, чтобы доказать некоторые важные утверждения о матрицах.

В главе 6 мы построили одностороннюю обратную матрицу для ортогональной матрицы

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix},$$

задающей переход от ортонормированного базиса $\{\mathbf{e}_i\}$ к ортонормированному базису $\{\mathbf{f}_i\}$. Столбцы матрицы перехода представляют собой выражения соответствующих векторов ортонормированного базиса $\{\mathbf{f}_i\}$, поэтому при $i \neq j$

$$c_{1i}c_{1j}+\ldots+c_{ni}c_{nj} = (\mathbf{f}_i, \mathbf{f}_j) = 0, \text{ а } c_{1i}c_{1i}+\ldots+c_{ni}c_{ni} = (\mathbf{f}_i, \mathbf{f}_i) = 1. \quad (8.1)$$

Это значит, как отмечалось в главе 6, что произведение $C'C$ равно единичной матрице E. Доказательство того факта, что CC' также единичная матрица и тем самым C' обратная матрица к C, мы отложили до настоящей главы. Восполним пробел.

Будем говорить в случае $AB = E$, что A является левой обратной к B, а B — правой обратной к A.

Теорема 8.1. *Если квадратная матрица A имеет левую обратную и правую обратную, то они совпадают.*

Доказательство. Пусть B левая, а C правая обратные к A. Рассмотрим произведение BAC. С одной стороны, $BAC = (BA)C = EC = C$, с другой стороны, $BAC = B(AC) = BE = B$. Тем самым $B = C$, а значит, $BA = AB = E$, и B (а равно и C) является двусторонней обратной матрицей к A. **Теорема доказана.**

Далее мы докажем, что ортогональная матрица C всегда обратима. Поскольку C^{-1}, если она существует, является, в частности, и правой обратной к C, то она совпадет с левой обратной C', т.е. для ортогональной матрицы $C' = C^{-1}$.

Итак, мы доказываем теорему об обратимости ортогональной матрицы, т.е. матрицы, столбцы которой удовлетворяют вышеприведенным равенствам (8.1).

Следующая теорема верна для любой, а не только для ортогональной матрицы.

Теорема 8.2. *Если столбцы матрицы C линейно независимы, то ее определитель не равен нулю.*

Доказательство. Предположим, что определитель матрицы равен нулю. Тогда система линейных уравнений

$$\begin{cases} c_{11}x_1 & + & c_{12}x_2 & + & \ldots & + & c_{1m}x_m & = & 0 \\ c_{21}x_1 & + & c_{22}x_2 & + & \ldots & + & c_{2m}x_m & = & 0 \\ & \ldots & & \ldots & & \ldots & & \ldots & \\ c_{n1}x_1 & + & c_{n2}x_2 & + & \ldots & + & c_{nm}x_m & = & 0 \end{cases},$$

имеет ненулевое решение $x_1 = \lambda_1, \ldots, x_n = \lambda_n$. Это означает, что

$$\lambda_1 \begin{pmatrix} c_{12} \\ c_{22} \\ \ldots \\ c_{n2} \end{pmatrix} + \ldots + \lambda_n \begin{pmatrix} c_{1n} \\ c_{2n} \\ \ldots \\ c_{nn} \end{pmatrix} = \mathbf{0}$$

и столбцы матрицы линейно зависимы, что противоречит условию теоремы. Тем самым **теорема доказана**.

Условие (8.1) ортогональности матрицы, вытекающее из ортогональности базиса $\{\mathbf{f_i}\}$, позволяет нам применить теорему 6.8, утверждающую, что система ненулевых взаимно ортогональных векторов линейно независима. По только что доказанной теореме это значит, что определитель ортогональной матрицы отличен от нуля, а это позволяет нам явно построить обратную к ней матрицу, как это делалось в четвертой главе.

Мы доказали, что обратной к ортогональной матрице является ее транспонированная. Из этого факта прямо следует, что C' также ортогональная матрица.

На языке матриц это объясняется тем, что $CC' = (C')'C' = E$, поэтому для столбцов C' также выполнено условие ортонормированности: при $i \neq j$

$$c'_{1i}c'_{1j} + \ldots + c'_{ni}c'_{nj} = 0, \text{ и } c'_{1i}c'_{1i} + \ldots + c'_{ni}c'_{ni} = 1.$$

На языке векторов мы можем объяснить это тем, что $C^{-1} = C'$ является матрицей обратного перехода от ортонормированного базиса $\{\mathbf{f_i}\}$ к ортонормированному базису $\{\mathbf{e_i}\}$, и ортонормированность столбцов C' вытекает из ортонормированности базиса $\{\mathbf{e_i}\}$.

Между прочим, мы доказали довольно странный с чисто алгебраической точки зрения факт: если столбцы матрицы ортонормированы, то также ортонормированы и ее строки.

Также не выглядит самоочевидным и простое следствие из теоремы 8.2: если столбцы матрицы линейно независимы, то также линейно независимы и ее строки. Это последнее утверждение доказывается следующим рассуждением. Если линейно независимы столбцы, то определитель матрицы не равен нулю, значит, не равен нулю и определитель транспонированной матрицы, значит, не имеет ненулевых решений система линейных уравнений, заданная транспонированной матрицей, значит, столбцы транспонированной матрицы линейно независимы, а они-то и суть строки исходной матрицы.

8.2. Билинейные формы

В главе 6 было дано следующее определение скалярного произведения в произвольном линейном пространстве:

Скалярным произведением векторов в линейном пространстве $\mathbf{V}$ может быть любая функция $\mathbf{S}(\mathbf{u},\mathbf{v})$, удовлетворяющая следующим условиям:

1) $\mathbf{S}(\mathbf{u},\mathbf{v}) = \mathbf{S}(\mathbf{v},\mathbf{u})$;
2) $\mathbf{S}(\mathbf{u}+\mathbf{u}',\mathbf{v}) = \mathbf{S}(\mathbf{u},\mathbf{v}) + \mathbf{S}(\mathbf{u}',\mathbf{v})$;
3) $\mathbf{S}(\lambda\mathbf{u},\mathbf{v}) = \lambda\mathbf{S}(\mathbf{u},\mathbf{v})$;
4) $\mathbf{S}(\mathbf{u},\mathbf{u}) > \mathbf{0}$.

➽ **Определение 8.5.** *Симметрической билинейной формой, заданной на элементах линейного пространства, называется функция, удовлетворяющая свойствам 1–3. Если дополнительно выполнено также и свойство 4, то симметрическая билинейная форма называется положительно определенной. Неотрицательно определенной назовем форму, для которой имеет место более слабое неравенство* $\mathbf{S}(\mathbf{u},\mathbf{u}) \geq \mathbf{0}$.

В пространстве, где нет естественных транспортиров и угольников, скалярное произведение задает ортогональность векторов и позволяет измерять иные углы между векторами. Как мы говорили, в произвольном линейном пространстве можно задать много разных скалярных произведений, и единственное, что требуется от такой функции двух векторных переменных, — это удовлетворять требованиям 1–4. Мы сейчас приступаем к рассмотрению таких функций.

В итоге это даст нам возможность рассматривать важное отношение между переменными — ковариацию — как аналог скалярного произведения, точнее даже, как второе скалярное произведение в пространстве исходных переменных (например, результатов тестирования), задающее в этом пространстве некоторую "прикладную" систему измерения углов.

8.3. Матрица билинейной формы

Многие определения и рассуждения этой главы могут быть легко распространены на более широкие классы функций и билинейных форм, однако для простоты далее под билинейными формами мы будем подразумевать только симметрические билинейные формы, вполне достаточные для наших целей.

Рассмотрим два вектора нашего линейного пространства и разложим их по ортонормированному базису $\{\mathbf{e_i}\}$: $\mathbf{x} = x_1\mathbf{e_1} + \ldots + x_n\mathbf{e_n}$, а $\mathbf{y} = y_1\mathbf{e_1} + \ldots + y_n\mathbf{e_n}$. Используя пункт 2 определения билинейной формы, получаем

$$\begin{gathered}\mathbf{S}(\mathbf{x},\mathbf{y}) = \\ = x_1y_1\mathbf{S}(\mathbf{e_1},\mathbf{e_1}) + \ldots + x_1y_n\mathbf{S}(\mathbf{e_1},\mathbf{e_n}) + \\ + \qquad \ldots \qquad + \\ + x_ny_1\mathbf{S}(\mathbf{e_n},\mathbf{e_1}) + \ldots + x_ny_n\mathbf{S}(\mathbf{e_n},\mathbf{e_n}).\end{gathered}$$

Зададим матрицу S билинейной формы $\mathbf{S}$ в данном базисе:

$$s_{ij} = \mathbf{S}(\mathbf{e_i},\mathbf{e_j}).$$

Это позволяет нам заменить в последней сумме $\mathbf{S}(\mathbf{e_i},\mathbf{e_j})$ на s_{ij}:

$$\begin{gathered}\mathbf{S}(\mathbf{x},\mathbf{y}) = \\ = x_1y_1s_{11} + \ldots + x_1y_ns_{1n} + \\ + \qquad \ldots \qquad + \\ + x_ny_1s_{n1} + \ldots + x_ny_ns_{nn}.\end{gathered}$$

Последнее выражение уже встречалось нам в шестой главе при доказательстве теоремы 4, утверждающей, что если линейное преобразование $\mathbf{A}$ имеет симметричную матрицу, то для любых двух векторов $\mathbf{u}$ и $\mathbf{v}$ имеет место равенство $(\mathbf{A}(\mathbf{u}),\mathbf{v}) = (\mathbf{u},\mathbf{A}(\mathbf{v}))$. Наше выражение не что иное, как скалярное произведение

$$\left(\begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}, \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1n} \\ s_{21} & s_{22} & \ldots & s_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ s_{n1} & s_{n2} & \ldots & s_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}\right).$$

Это значит, что самосопряженное линейное преобразование, имеющее ту же матрицу S, может быть использовано для вычисления значений билинейной формы. Обозначим это преобразование $\mathbf{A_S}$.

Соотношение

$$\mathbf{S}(\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{A_S}(\mathbf{y}))$$

выполняется для векторов нашего пространства $\mathbf{V}$ независимо от выбора базиса.

8.4. Главные оси билинейной формы

Матрица билинейной формы совпадает с матрицей соответствующего линейного преобразования и поэтому преобразуется при замене ортонормированных базисов по той же самой формуле $S_f = C'S_eC$.

Если мы найдем собственные значения и собственные векторы матрицы S, мы тем самым найдем собственный базис для преобразования $\mathbf{A_S}$. В этом собственном базисе $\{\mathbf{f_i}\}$ выражение для билинейной формы

$$\mathbf{S}(\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{A_S}(\mathbf{y}))$$

запишется наиболее просто:

$$\left(\begin{pmatrix} x_1' \\ x_2' \\ \dots \\ x_n' \end{pmatrix}, \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \begin{pmatrix} y_1' \\ y_2' \\ \dots \\ y_n' \end{pmatrix}\right).$$

В этом случае говорят, что билинейная форма приведена к главным осям. Главное преимущество, которое дают нам главные оси, это то, что $\mathbf{S}(\mathbf{f_i},\mathbf{f_j}) = 0$ при $i \neq j$. Если билинейная форма положительно определена, то она может быть использована как второе скалярное произведение в линейном пространстве $\mathbf{V}$. Соотношение $\mathbf{S}(\mathbf{f_i},\mathbf{f_j}) = 0$ интерпретируется тогда следующим образом: базис $\{\mathbf{f_i},\mathbf{f_j}\}$ ортонормирован в смысле исходного скалярного произведения и ортогонален в смысле второго скалярного произведения.

8.5. Матрица выборочной ковариации

Наша следующая цель — доказать, что матрица выборочной ковариации, о которой мы начали говорить в предыдущей главе, преобразуется при замене базиса по тому же закону, что и матрица линейного преобразования, а значит, может быть приведена к диагональному виду.

Пусть даны k векторов в n-мерном пространстве:

$$\begin{pmatrix} x_1^{(1)} \\ \dots \\ x_n^{(1)} \end{pmatrix}; \quad \begin{pmatrix} x_1^{(2)} \\ \dots \\ x_n^{(2)} \end{pmatrix}; \quad \dots; \begin{pmatrix} x_1^{(k)} \\ \dots \\ x_n^{(k)} \end{pmatrix}.$$

Здесь нам полезно будет иметь в виду обычно подразумеваемый в психологических исследованиях смысл этих векторов: например, каждый вектор представляет собой набор тестовых показателей по n тестам испытуемого, номер которого задается верхним, помещенным в скобки индексом.

Матрица средних попарных зависимостей между координатами векторов вычисляется следующим образом: на место ij помещается среднее по k (испытуемым) значение произведения i-й и j-й координат $r_{ij} = \frac{1}{k}(x_i^{(1)}x_j^{(1)} + x_i^{(2)}x_j^{(2)} + \ldots + x_i^{(k)}x_j^{(k)})$.

На i-м месте главной диагонали будет стоять тогда средняя сумма квадратов $r_{ii} = \frac{1}{k}((x_i^{(1)})^2 + (x_i^{(2)})^2 + \ldots + (x_i^{(k)})^2)$.

Полученная матрица

$$\frac{1}{k}\begin{pmatrix} (x_1^{(1)})^2+(x_1^{(2)})^2+\ldots+(x_1^{(n)})^2 & \ldots x_1^{(1)}x_n^{(1)}+x_1^{(2)}x_n^{(2)}+\ldots+x_1^{(k)}x_n^{(k)} \\ x_2^{(1)}x_1^{(1)}+x_2^{(2)}x_1^{(2)}+\ldots+x_2^{(k)}x_1^{(k)} & \ldots x_2^{(1)}x_n^{(1)}+x_2^{(2)}x_n^{(2)}+\ldots+x_2^{(k)}x_n^{(k)} \\ \ldots & \ldots \qquad \ldots \\ x_n^{(1)}x_1^{(1)}+x_n^{(2)}x_1^{(2)}+\ldots+x_n^{(k)}x_1^{(k)} & \ldots \quad (x_n^{(1)})^2+(x_n^{(2)})^2+\ldots+(x_n^{(k)})^2 \end{pmatrix}$$

симметрична, поскольку $r_{ij} = r_{ji}$. Обозначим нашу матрицу через R.

Теорема 8.3. *Если C матрица перехода от исходного базиса к новому, то матрица R преобразуется по формуле $R^{new} = C'RC$.*

Доказательство. Разложим матрицу R в сумму матриц попарных зависимостей, вычисленных для каждого испытуемого отдельно:
$R = \frac{1}{k}(R^{(1)} + R^{(2)} + \ldots + R^{(k)})$, где

$$R^{(k)} = \begin{pmatrix} (x_1^{(k)})^2 & x_1^{(k)}x_2^{(k)} & \ldots & x_1^{(k)}x_n^{(k)} \\ x_2^{(k)}x_1^{(k)} & (x_2^{(k)})^2 & \ldots & x_2^{(k)}x_n^{(k)} \\ \ldots & \ldots & \ldots & \\ x_n^{(k)}x_1^{(k)} & x_n^{(k)}x_2^{(k)} & \ldots & (x_n^{(k)})^2 \end{pmatrix}.$$

Если мы докажем, что при замене базиса, заданном матрицей перехода C, каждое из слагаемых $R^{(k)}$ преобразуется по закону $R^{(k)}_{new} = C'R^{(k)}C$, то и для суммы имеет место равенство $R_{new} = C'RC$, поскольку тогда

$$\begin{aligned} R_{new} &= \tfrac{1}{k}(R^{(1)}_{new} + R^{(2)}_{new} + \ldots + R^{(k)}_{new}) = \\ &= \tfrac{1}{k}(C'R^{(1)}C + C'R^{(2)}C + \ldots + C'R^{(k)}C) = \\ &= \tfrac{1}{k}C'(R^{(1)} + R^{(2)} + \ldots + R^{(k)})C = \\ &= C'R^{(k)}C. \end{aligned}$$

Докажем, что $R^{(k)}_{new} = C'R^{(k)}C$. Опуская для краткости индексы (k), предположим, что $\begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix}$ результаты одного испытуемого, а

$$R = \begin{pmatrix} (x_1)^2 & x_1x_2 & \ldots & x_1x_n \\ x_2x_1 & (x_2)^2 & \ldots & x_2x_n \\ \ldots & \ldots & \ldots & \\ x_nx_1 & x_nx_2 & \ldots & (x_n)^2 \end{pmatrix}$$ соответствующая матрица.

Заметим теперь, что эту матрицу можно получить, перемножив как матрицы вектор-столбец на вектор-строку:

$$R = \begin{pmatrix} (x_1)^2 & x_1x_2 & \ldots & x_1x_n \\ x_2x_1 & (x_2)^2 & \ldots & x_2x_n \\ \ldots & \ldots & \ldots & \\ x_nx_1 & x_nx_2 & \ldots & (x_n)^2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} (x_1x_2 \ldots x_n).$$

Эта непривычная форма умножения позволит нам коротко получить нужный результат.

Пусть матрица перехода C связывает старые координаты (в базисе вопросов) с новыми (скажем, в базисе факторов) формулой

$$\begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix} = \begin{pmatrix} c_{11} & \ldots & c_{1n} \\ \ldots & \ldots & \ldots \\ c_{n1} & \ldots & c_{nn} \end{pmatrix} \begin{pmatrix} x'_1 \\ \ldots \\ x'_n \end{pmatrix}.$$

Это, как мы знаем, означает, что новые координаты через старые выражаются посредством матрицы, обратной к C, а поскольку базисы наши ортонормированные, то для транспонированной C':

$$\begin{pmatrix} x'_1 \\ \ldots \\ x'_n \end{pmatrix} = \begin{pmatrix} c_{11} & \ldots & c_{n1} \\ \ldots & \ldots & \ldots \\ c_{1n} & \ldots & c_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix}.$$

Возьмем теперь модификацию этого равенства:

$$\begin{pmatrix} x_1' & \dots & x_n' \end{pmatrix} = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} c_{11} & \dots & c_{1n} \\ \dots & \dots & \dots \\ c_{n1} & \dots & c_{nn} \end{pmatrix}.$$

Это последнее равенство можно непосредственно сопоставить с предыдущим и убедиться, что формулы пересчета идентичны. Но можно также увидеть, что если первое равенство записать в матричном виде как $K = MN$, то второе выражает равенство транспонированных матриц: $K' = (NM)'$ — надо только увидеть, что транспонированный столбец (матрица $1 \times n$) это строка (матрица $n \times 1$).

Научившись видеть в столбцах и строках полноправные матрицы, запишем теперь произведение четырех матриц

$$\begin{pmatrix} c_{11} & \dots & c_{n1} \\ \dots & \dots & \dots \\ c_{1n} & \dots & c_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix} \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} c_{11} & \dots & c_{1n} \\ \dots & \dots & \dots \\ c_{n1} & \dots & c_{nn} \end{pmatrix}.$$

Мы можем преобразовать это произведение двумя способами.

1) Воспользовавшись предыдущими двумя равенствами, заменим первую пару матриц на столбец $\begin{pmatrix} x_1' \\ \dots \\ x_n' \end{pmatrix}$, а вторую на строку $\begin{pmatrix} x_1' \dots x_n' \end{pmatrix}$.

В результате получим уже встречавшееся произведение столбца на строку

$$\begin{pmatrix} x_1' \\ x_2' \\ \dots \\ x_n' \end{pmatrix} (x_1' x_2' \dots x_n') = \begin{pmatrix} (x_1')^2 & x_1'x_2' & \dots & x_1'x_n' \\ x_2'x_1' & (x_2')^2 & \dots & x_2'x_n' \\ \dots & \dots & \dots & \\ x_n'x_1' & x_n'x_2' & \dots & (x_n')^2 \end{pmatrix},$$

а это и есть матрица R^{new}.

2) Центральная пара матриц — произведение столбца на строку — есть исходная матрица R:

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} (x_1 x_2 \dots x_n) = \begin{pmatrix} (x_1)^2 & x_1x_2 & \dots & x_1x_n \\ x_2x_1 & (x_2)^2 & \dots & x_2x_n \\ \dots & \dots & \dots & \\ x_nx_1 & x_nx_2 & \dots & (x_n)^2 \end{pmatrix} = R,$$

а значит, все произведение четырех матриц может быть представлено как

$$C'RC.$$

Таким образом, преобразуя одну и ту же формулу, мы в первом случае получили матрицу R^{new}, а во втором $C'RC$. Это означает, что

$$R^{new} = C'RC,$$

и теорема доказана.

На практике матрицы такого типа, как мы рассмотрели выше, вычисляются не для произвольных выборок тестовых результатов, а для центрированных и чаще всего стандартизованных[1] выборок.

Если для любого i (т.е. для любой переменной) выполнено условие $x_i^{(1)} + x_i^{(2)} + \ldots + x_i^{(k)} = 0$, то выборка называется центрированной. Для того чтобы центрировать выборку по данной переменной, надо вычесть из каждого значения среднее арифметическое по выборке, используя формулу

$$x_i'^{(k)} = x_i^{(k)} - (x_i^{(1)} + x_i^{(2)} + \ldots + x_i^{(k)})/k.$$

Упражнение 8.1. Проверить, что $x_i'^{(1)} + x_i'^{(2)} + \ldots + x_i'^{(k)} = 0$.

Упражнение 8.2. Проверить, что это преобразование не является линейным в смысле определения главы 4.

Можно представить это преобразование в векторном виде: как и прежде имеется выборка — k векторов в n-мерном пространстве:

$$v^{(1)} = \begin{pmatrix} x_1^{(1)} \\ \ldots \\ x_n^{(1)} \end{pmatrix}; \quad v^{(2)} = \begin{pmatrix} x_1^{(2)} \\ \ldots \\ x_n^{(2)} \end{pmatrix}; \quad \ldots; v^{(k)} = \begin{pmatrix} x_1^{(k)} \\ \ldots \\ x_n^{(k)} \end{pmatrix}.$$

Положим

$$\bar{v} = \frac{v^{(1)} + v^{(2)} + \ldots + v^{(k)}}{k}.$$

Каждая компонента вектора $\bar{v}$ будет средним арифметическим соответствующих компонент векторов $v^{(1)}, v^{(2)}, \ldots, v^{(k)}$.

Положим далее

$$v'^{(1)} = v^{(1)} - \bar{v}, \quad v'^{(2)} = v^{(2)} - \bar{v}, \quad \ldots, v'^{(k)} = v^{(k)} - \bar{v}.$$

[1] Общее определение будет дано несколько позже.

Упражнение 8.3. Проверить, что

$$v'^{(1)} + v'^{(2)} + \ldots + v'^{(k)} = 0.$$

Заметим, что после того, как центрирование проведено, последнее соотношение будет выполнено в любом базисе, т.е. после замены базиса суммы по каждой координате всех k векторов всегда останутся нулевыми.

➽ Определение 8.6. *Если выборка центрирована по всем переменным, то матрица попарных зависимостей R называется матрицей ковариаций.*

Теорема 8.4. *Матрица выборочных ковариаций задает неотрицательно определенную билинейную форму.*

Доказательство. Имеем в исходном базисе $\{\mathbf{e_i}\}$ матрицу ковариаций

$$R_e =$$

$$= \frac{1}{k}\begin{pmatrix} (x_1^{(1)})^2 + (x_1^{(2)})^2 + \ldots + (x_1^{(n)})^2 & \ldots x_1^{(1)} x_n^{(1)} + x_1^{(2)} x_n^{(2)} + \ldots + x_1^{(k)} x_n^{(k)} \\ x_2^{(1)} x_1^{(1)} + x_2^{(2)} x_1^{(2)} + \ldots + x_2^{(k)} x_1^{(k)} & \ldots x_2^{(1)} x_n^{(1)} + x_2^{(2)} x_n^{(2)} + \ldots + x_2^{(k)} x_n^{(k)} \\ \ldots & \ldots \quad\quad \ldots \\ x_n^{(1)} x_1^{(1)} + x_n^{(2)} x_1^{(2)} + \ldots + x_n^{(k)} x_1^{(k)} & \ldots \quad (x_n^{(1)})^2 + (x_n^{(2)})^2 + \ldots + (x_n^{(k)})^2 \end{pmatrix}.$$

Воспользовавшись разложениями векторов линейного пространства $\mathbf{V}$ по данному базису и матрицей R_e, определим билинейную форму формулой $\mathbf{R}(\mathbf{u}, \mathbf{v}) = (u_e, R_e v_e)$.

Надо доказать, что для любого вектора $\mathbf{v}$ неотрицательна величина $\mathbf{R}(\mathbf{v}, \mathbf{v})$.

Как показано в теореме 3, в любом базисе $\{\mathbf{f_i}\}$ на главной диагонали матрицы R_f будут стоять суммы квадратов координат векторов выборки, пересчитанных в новом базисе:

$$r_{ii} = (x_i'^{(1)})^2 + (x_i'^{(2)})^2 + \ldots + (x_i'^{(n)})^2.$$

Это значит, что элементы главной диагонали матрицы ковариаций всегда неотрицательны.

Возьмем состоящий из собственных векторов формы $\mathbf{R}$ ортонормированный базис $\{\mathbf{f_i}\}$. Поскольку базис $\{\mathbf{f_i}\}$ собственный, вне главной

диагонали в матрице стоят нули, а диагональные элементы неотрицательны.

Разложим вектор v по базису $\{\mathbf{f_i}\}$ и вычислим $\mathbf{R}(\mathbf{v},\mathbf{v})$ по этому разложению и матрице R_f.

$$\mathbf{R}(\mathbf{v},\mathbf{v}) =$$

$$= \left(\begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix}, \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix} \right) =$$

$$= \lambda_1(v_1)^2 + \lambda_2(v_2)^2 + \ldots + \lambda_n(v_n)^2.$$

Последнее выражение не может быть меньше нуля, так как $\lambda_i \geq 0$. **Теорема доказана.**

8.6. Матрица корреляции

В предыдущей главе мы выяснили, что при отсутствии веских соображений об уместности определенных единиц измерения имеет смысл стандартизовать переменные.

Пусть дана центрированная выборка

$$v^{(1)} = \begin{pmatrix} x_1^{(1)} \\ \dots \\ x_n^{(1)} \end{pmatrix}; \ v^{(2)} = \begin{pmatrix} x_1^{(2)} \\ \dots \\ x_n^{(2)} \end{pmatrix}; \ \dots; \ v^{(k)} = \begin{pmatrix} x_1^{(k)} \\ \dots \\ x_n^{(k)} \end{pmatrix}.$$

➽ Определение 8.7. *Дисперсия переменной x_i по данной выборке вычисляется по формуле*

$$D_i = \frac{1}{k}(x_i^{(1)})^2 + (x_i^{(2)})^2 + \ldots + (x_i^{(k)})^2.$$

➽ Определение 8.8. *Выборка называется стандартизованной, если ее дисперсии по каждой переменной равны единице.*

Всякую выборку можно стандартизовать. Проверим, что следующая выборка является стандартизованной:

$$v'^{(1)} = \begin{pmatrix} x_1^{(1)}/\sqrt{D_1} \\ \dots \\ x_n^{(1)}/\sqrt{D_n} \end{pmatrix}; \ v'^{(2)} = \begin{pmatrix} x_1^{(2)}/\sqrt{D_1} \\ \dots \\ x_n^{(2)}/\sqrt{D_n} \end{pmatrix}; \ \dots; \ v'^{(k)} = \begin{pmatrix} x_1^{(k)}/\sqrt{D_1} \\ \dots \\ x_n^{(k)}/\sqrt{D_n} \end{pmatrix}.$$

Действительно, посчитаем дисперсию переменной с номером i по новой выборке:

$$\frac{1}{k}((x_i^{(1)}/\sqrt{D_i})^2 + (x_i^{(2)}/\sqrt{D_i})^2 + \dots + (x_i^{(k)}/\sqrt{D_i})^2) =$$

$$= \frac{1}{k}((x_i^{(1)})^2 + (x_i^{(2)})^2 + \dots + (x_i^{(k)})^2)/D = D/D = 1.$$

Таким образом, наша выборка стандартизована.

Определение 8.9. *Суммарная дисперсия выборки равна сумме дисперсий переменных:* $D = D_1 + D_2 + \dots + D_n$.

Суммарная дисперсия стандартизованной выборки равна количеству переменных, поскольку каждая из них имеет дисперсию равную единице.

Теорема 8.5. *Суммарная дисперсия выборки не зависит от базиса.*

Доказательство. Докажем для упрощения вычислений, что не зависит от базиса произведение Dk. Запишем по строкам $D_1k + D_2k + \dots + D_nk$.

$$Dk =$$
$$= (x_1^{(1)})^2 + (x_1^{(2)})^2 + \dots + (x_1^{(k)})^2 +$$
$$+ (x_2^{(1)})^2 + (x_2^{(2)})^2 + \dots + (x_2^{(k)})^2 +$$
$$\dots$$
$$+ (x_n^{(1)})^2 + (x_n^{(2)})^2 + \dots + (x_n^{(k)})^2.$$

Теперь заметим, что по столбцам записаны скалярные квадраты векторов $v^{(1)}, v^{(2)}, \dots v^{(k)}$ и, следовательно, суммарная дисперсия, умноженная на k, есть сумма скалярных квадратов векторов, составляющих выборку, а скалярные произведения не меняются при замене базиса. **Теорема доказана.**

Когда матрица ковариаций (или корреляций — в данном случае это безразлично) приведена к диагональному виду, суммарная дисперсия равна сумме собственных значений $\lambda_1 + \ldots + \lambda_n$.

➡ **Определение 8.10.** *Долей общей дисперсии выборки, приходящейся на i-й фактор, называется отношение*

$$\frac{\lambda_i}{\lambda_1 + \ldots + \lambda_n}.$$

Естественно, что сумма долей дисперсии по всем факторам равна единице.

Собственные векторы билинейной формы $\mathbf{R}(\mathbf{u}, \mathbf{v})$, заданной в исходном базисе матрицей выборочных корреляций исходных переменных, интерпретируются обычно как скрытые независмые факторы, характеризующие исследуемый предмет. Их нельзя наблюдать непосредственно, но по вектору, характеризующему данный элемент выборки (по первичным результатам данного испытуемого) их значения можно восстановить с помощью матрицы перехода.

Напомним, что у матрицы перехода к собственному базису по столбцам расположены разложения по исходному базису именно собственных векторов — элементов нового базиса.

Запишем выражение новых координат через старые и матрицу перехода:

$$v_f = C' v_e,$$

или подробнее для данного испытуемого (опускаем верхний индекс k)

$$\begin{pmatrix} v_1' \\ v_2' \\ \ldots \\ v_n' \end{pmatrix} = \begin{pmatrix} c_{11} & c_{21} & \ldots & c_{n1} \\ c_{12} & c_{22} & \ldots & c_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ c_{1n} & c_{2n} & \ldots & c_{nn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \ldots \\ v_n \end{pmatrix},$$

Координаты в базисе факторов v_i' называются факторными значениями испытуемого по i-му фактору.

Выборка в координатах собственного базиса не является стандартизованной. Дисперсии соответствующих факторам переменных (факторных значений) равны собственным значениям матрицы корреляций.

8.7. Углы между исходными переменными и факторами. Факторные нагрузки

Напомним, что вектор старого базиса $\mathbf{e_i}$, задающий исходную переменную, разлагается по базису факторов по формуле

$$\mathbf{e_i} = c_{i1}\mathbf{f_1} + c_{i2}\mathbf{f_2} + \ldots + c_{in}\mathbf{f_n}.$$

(Выражение векторов нового базиса через старый записывалось по столбцам матрицы перехода, а старого через новый — по строкам.) Умножим скалярно обе части равенства на $\mathbf{f_1}$

$$(\mathbf{e_i}, \mathbf{f_1}) = c_{i1}(\mathbf{f_1}, \mathbf{f_1}) + c_{i2}(\mathbf{f_2}, \mathbf{f_1}) + \ldots + c_{in}(\mathbf{f_n}, \mathbf{f_1})$$

и заметим, что в правой части только первое скалярное проиведение равно единице, а остальные равны нулю в силу ортогональности базиса $\{\mathbf{f_i}\}$, поэтому

$$(\mathbf{e_i}, \mathbf{f_1}) = c_{i1}.$$

Аналогично

$$(\mathbf{e_i}, \mathbf{f_j}) = c_{ij},$$

а поскольку векторы базиса имеют единичную длину, то по аналогии со скалярным произведением векторов в пространстве, в котором мы живем, c_{ij} можно считать косинусом угла между соответствующими исходной переменной и фактором.

Однако в факторном анализе более важны углы, расчитанные по альтернативному скалярному произведению, заданному билинейной формой $\mathbf{R}(\mathbf{x}, \mathbf{y})$, порожденной матрицей корреляций.

Предположим, что данная форма положительно определена, т.е. $\mathbf{R}(\mathbf{x}, \mathbf{x}) > \mathbf{0}$, что в приложениях бывает практически всегда (это значит, что все собственные значения формы строго больше нуля, хотя и могут быть очень маленькими числами).

Для большей наглядности обозначим второе скалярное произведение двойными скобками:

$$((\mathbf{x}, \mathbf{y})) = \mathbf{R}(\mathbf{x}, \mathbf{y}).$$

Запишем в базисе факторов новое скалярное произведение i-й исходной переменной и j-го фактора. Для этого выразим i-й вектор старого базиса через факторы и матрицу перехода:

$$\mathbf{e_i} = c_{i1}\mathbf{f_1} + c_{i2}\mathbf{f_2} + \ldots + c_{in}\mathbf{f_n}.$$

Умножим, как и в предыдущем случае, обе части равенства на $\mathbf{f_j}$, только используя второе скалярное произведение. Заметим, что векторы $\mathbf{f_i}$ ортогональны и в смысле второго произведения, поэтому, как и в первом случае, пропадут все слагаемые, кроме одного:

$$((\mathbf{e_i}, \mathbf{f_j})) = c_{ij}((\mathbf{f_j}, \mathbf{f_j})).$$

Однако $((\mathbf{f_j}, \mathbf{f_j})) \neq 1$ в отличие от $(\mathbf{f_j}, \mathbf{f_j})$.

Вспомним теперь, что $((\mathbf{f_j}, \mathbf{f_j})) = \mathbf{R}(\mathbf{f_j}, \mathbf{f_j})$. Для того чтобы посчитать этот новый скалярный квадрат, выразим $\mathbf{f_j}$ и матрицу R в собственном базисе, элементом которого f_j является.

$$\mathbf{R}(\mathbf{f_j}, \mathbf{f_j}) = \left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_j & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & \lambda_n \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{pmatrix} \right) = \lambda_{\mathrm{j}}.$$

Таким образом, $((\mathbf{e_i}, \mathbf{f_j})) = c_{ij}\lambda_j$.

По аналогии с обычным скалярным произведением, для того чтобы найти косинус угла между i-й переменной и j-м фактором, надо скалярное произведение поделить на модули сомножителей — в нашем случае на корни из новых скалярных квадратов сомножителей:

$$\cos\alpha_{ij} = \frac{((\mathbf{e_i}, \mathbf{f_j}))}{\sqrt{((\mathbf{e_i}, \mathbf{e_i}))}\sqrt{((\mathbf{f_i}, \mathbf{f_i}))}}.$$

К счастью, мы уже нашли, что $((\mathbf{f_i}, \mathbf{f_i})) = \lambda_j$. Осталось вычислить $((\mathbf{e_i}, \mathbf{e_i})) = \mathbf{R}(\mathbf{e_i}, \mathbf{e_i})$.

Нам опять "повезло": $\mathbf{R}(\mathbf{e_i}, \mathbf{e_i}) = \mathbf{1}$, поскольку наша выборка стандартизована как раз по отношению к переменным старого базиса.

Окончательно получаем

$$\cos\alpha_{ij} = \frac{c_{ij}\lambda_j}{\sqrt{\lambda_j}} = c_{ij}\sqrt{\lambda_j}.$$

Последнее выражение называется факторной нагрузкой i-й переменной на j-фактор.

Часть II

Математический анализ

Глава 1

Исходные идеи дифференциального исчисления

1.1. Историко-философский экскурс

Проблемы, встающие перед изучающим математический анализ, существенно отличаются от трудностей изучения линейной алгебры. В нескольких словах: различие определяется тем, что линейная алгебра имеет дело с объектами в определенном смысле конечными и позитивными, а анализ — с бесконечными и парадоксальными. Мы здесь не будем пытаться исчерпывающим образом уточнять сказанное — это потребовало бы еще одной книги, — однако и скрывать данное обстоятельство от читателя, тем более от психолога, представляется большой ошибкой. На наш взгляд, люди, среди предметов профессионального интереса которых мышление, язык, знаки, немало приобретут от знакомства с той сферой мышления, которая именуется "математический анализ".

Прежде всего, отметим связь между понятийными трудностями анализа и известными парадоксами Зенона, исследующими вопрос, как можно мыслить "микроструктуру" пространства и времени: что такое момент времени? можно ли понимать время как последовательность моментов? как понимать высказывание: движущийся объект A в настоящий момент находится в точке X?

Современный способ изложения математического анализа вызревал в течение более чем двух столетий, и наиболее важные сдвиги происходили не в решении прикладных задач (огромное их число умели решать в XVII и XVIII веках), а именно в основаниях анализа и в его понятийном аппарате. Первоначальные формулировки идей анализа, принадлежавшие, как принято считать, Ньютону и Лейбницу[1], совершенно не похожи на те, которые мы можем найти в университетском учебнике сегодня.

Идеи творцов анализа были просты, но неясны, поэтому критики находили весьма серьезные аргументы против их построений. Во многом благодаря этой критике математики уточняли понятия анализа, что и привело в конце концов к современному состоянию. Хотя, как признают многие, не все в обосновании анализа гладко и в сегодняшней его версии; в учебнике мы не будем это обсуждать, согласившись, что по сравнению с временами Ньютона и Лейбница достигнут значительный прогресс. Однако ставшее ныне традиционным изложение анализа начиная с оснований представляется нам оправданным далеко не во всех случаях. Анализ для психологов именно такой случай, когда изложение может и должно быть иным.

Как и в линейной алгебре, изложение расслоится на линию идей и примеров и линию обоснований и доказательств. Первая линия будет выдержана в эклектическом стиле, включая как современные идеи, так и исходные идеи творцов анализа, которые редко посещают учебники. Вторая же линия будет целиком современной. Предполагается, что продумывая первую линию, читатель приобретет, во-первых, относительно небольшую по объему, осмысленную систему понимания анализа, которая позволит ему разумно употреблять понятия интеграла и производной в дальнейшем. Во-вторых, сама неясность оснований этой версии анализа даст возможность читателю понять необходимость того движения математического языка, которое привело анализ к его нынешней систематизации.

* * *

Расчет площадей и объемов был достаточно распространенной задачей, начиная с античных времен. Задачи эти назывались квадратура-

[1] Есть иные мнения. Например, в работах В.И. Арнольда приоритет отдается Гуку и Гюйгенсу. Проблемы приоритетов мы не будем здесь касаться. Нет сомнений, что у творцов анализа были учителя и предшественники, ученики и последователи. Важнее всего то, что их общими усилиями в XVI и XVII веках была произведена глубокая революция в математике и науке вообще.

ми и кубатурами, что хорошо проясняет их суть: площадь квадрата и объем куба считаются интуитивно ясными предметами, более трудные случаи следует сводить к этим известным.

Чисто геометрическими средствами, пользуясь только циркулем и линейкой, легко построить квадрат, площадь которого равна площади данного прямоугольника. Не труднее свести площадь треугольника к площади прямоугольника.

Принципиально более трудная задача — квадратура круга. Как оказалось, эту задачу нельзя решить точно[2], можно только приближаться к точному решению, расчитывая площади вписанных и описанных многоугольников с возрастающим количеством сторон. Легко видеть, что площадь круга будет всегда находиться между сходящимися к ней верхней оценкой (описанный многоугольник содержит круг) и нижней оценкой (вписанный многоугольник содержится в круге). Принципиальная трудность, отличающая эту задачу от задачи вычисления площади прямоугольника, состоит не в том, что мы не можем точно узнать площадь данного круга (в практических задачах и площадь данного прямоугольника мы не можем узнать точно), а в том, что само геометрическое определение площади круга можно дать только через так или иначе понимаемый предел последовательности площадей. В то же время здравый смысл говорит, что площадь круга вполне корректное понятие: например, практически ее можно определить через вес картонного круга данного размера, деленный на вес единичного квадрата из этого же картона. Здравый смысл будет утверждать, что наличие предела последовательности площадей многоугольников гарантируется этой “картонной” моделью понятия площади круга. Точно так же обстоит дело с объемами сосудов. Практически объем бочки легко определить, вычерпывая ее содержимое мерным сосудом, но геометрически его можно определить (заметьте различие значений слова “определить”) только с помощью какой-то бесконечной процедуры.

Интересный вопрос для психолога, культуролога и философа: почему в развитой математике античных греков не появилось дифференциальное исчисление? Объясняется этот факт тем, что бесконечность сама по себе ни в каких своих ипостасях не привлекала античных мыслителей, а скорее отпугивала. Греки были людьми конечной формы и окончательной истины, бесконечность была для них всегда “дурной”, “бесформенной”. Совершенно другая идеологическая ситуация сложи-

---

[2] Нельзя построить с помощью циркуля и линейки квадрат, площадь которого равна площади круга данного радиуса.

лась в XVII веке — христианская культура того времени относилась к бесконечности совершенно иначе.

Было бы ошибкой характеризовать возникновение дифференциального исчисления как теоретический ответ на практический запрос. Скорее, в основе научной революции XVII века мы находим неразделимое стремление к освоению мира (в том числе и практическому освоению) через познание замысла Творца. Творец бесконечно превосходит человека, и бесконечность мыслится как символ этого превосходства, и сама по себе вызывает даже теоретический интерес. Как видим, этот пиетет по отношению к бесконечному, который вполне отчетливо чувствуется и в современной культуре, — достаточно позднее приобретение.

Парадоксы Зенона предостерегали грека от прикосновения к бесконечно малым частям времени и пространства. Для пытливого европейца Нового времени бесконечно малые представляли собой вызов, требующий ответа.

На рубеже XVI и XVII веков почти одновременно были описаны законы движения планет (законы Кеплера) и законы свободного падения тел под действием земного тяготения (законы Галилея). Напомним, что законы Кеплера говорили, что всякая планета движется вокруг Солнца по эллиптической орбите, причем площади эллиптических секторов, проходимых ею в единицу времени, равны в любой точке орбиты. Законы Галилея описывали квадратичную зависимость между временем падения и проходимым расстоянием. Эти законы были эмпирическими, т.е. обобщали полученные наблюдения и эксперименты.

В последующие годы эти законы удалось объединить в едином ньютоновском понимании тяготения, пространства и времени. Чтобы стало понятно, как математики того времени могли относиться к этим успехам, приведем понятную недавним школьникам аналогию.

Решая задачу из сборника задач, школьник не без оснований считает подтверждением правильности своего решения тот факт, что в последнем уравнении под знаком корня после сложения 201 и 88 появляется 289 и корень удается извлечь — он равен 17. Примерно так же чувствовали себя Ньютон и его коллеги, когда аппарат дифференциального исчисления, который ими создавался, позволял получить и законы Кеплера, и законы Галилея точным выводом из весьма простых теоретических предпосылок. Легко догадаться, кого они считали автором "задачника природы", в котором задачи были составлены с теми же метками успеха, как и в современных школьных задачниках. Легко понять, участниками какой захватывающей игры они себя представляли.

* * *

Привяжем на полуметровую веревку небольшой грузик и будем вращать его над головой. Грузик движется по кругу. Какова его скорость в каждый момент? С одной стороны, если мы внезапно отпустим веревку, то грузик полетит по касательной со скоростью, как можно предположить, которую он имел в тот самый момент, когда был отпущен. Таким образом, мгновенная скорость в момент отпускания веревки являет себя в полете отпущенного груза. Предположим по аналогии, что и в другие моменты вращения мгновенная скорость вращающегося грузика в данной точке A направлена по касательной, проведенной к окружности в точке A. Сколько времени летит грузик с этой скоростью.

Парадокс I. *Если продолжительность полета с данной мгновенной скоростью не равна нулю, то, двигаясь по касательной, грузик покинет окружность. Если продолжительность полета с данной мгновенной скоростью равна нулю, то либо (1) грузик никуда не сдвинется, либо (2) движение должно получаться суммированием бесконечного числа нулевых сдвигов.*

Парадоксальным вариантом (1) довольствовался бы Зенон, вариант (2) выбрали математики XVII века. Они занялись суммированием бесконечно малых величин. Мы последуем за ними и сначала определим мгновенную скорость.

➽ **Почти определение 1**. *Мгновенная скорость грузика (будем считать его точечным) в данной точке траектории это предельное отношение пройденного грузиком пути, начиная от данной точки, к времени прохождения этого пути, если это время бесконечно приближается к нулю.*

Пример 1. В простейшем случае ничего парадоксального в этом определении не обнаруживается. Предположим, что грузик движется по прямой с постоянной скоростью v. В момент t грузик находится в некоторой точке x, в момент $t+o$ — в точке $x+vo$. Искомое отношение при любом значении o, как бы ни было оно близко к нулю, таково: $vo/o = v$. Мгновенная скорость равна постоянной скорости перемещения[3]

[3] Мы обозначили приращение времени буквой o, как это делали во времена Ньютона, чтобы подчеркнуть ее "родство" с нулем и облегчить тем самым восприятие символа с совершенно особой функцией. В современной литературе обычно используется обозначение Δx.

Пример 2. Пусть грузик падает вертикально под действием силы тяжести. Из опытов Галилея мы знаем две формулы, связывающих скорость v, путь S, время t и постоянное ускорение свободного падения g: $v = gt$ и $S = gt^2/2$. Для простоты здесь и далее будем считать, что ускорение свободного падения g равно не 9,8 м/сек^2, а 1 м/сек^2, каково оно, например, на высоте около 15 000 км от поверхности Земли.

В таком случае расстояние грузика от точки начала падения в момент времени t равно $t^2/2$, а через время o в момент $(t + o)$ равно $(t + o)^2/2$. За время o он проходит

$$(t + o)^2/2 - t^2/2 = (t^2 + 2to + o^2 - t^2)/2 = (to + o^2/2).$$

Отношение пути к времени, т.е. средняя скорость за время o, равно

$$v(o) = (to + o^2/2)/o = t + o/2.$$

От первой галилеевской, учитывая, что $g = 1$, полученную нами формулу отличает только слагаемое $o/2$. В нашем почти определении 1 говорится о предельном отношении при o неограниченно приближающемся к нулю. Заметим, что чем меньше становится o, тем меньше величина расхождения $o/2$, тем ближе наша формула к галилеевской. В последний момент перед "исчезновением" o, отношение S/o становится в точности равным галилеевскому. Как понимать этот последний момент перед исчезновением? Во времена Ньютона внятного ответа на этот вопрос не давалось. Смысл операции ясен: наше отношение "стремится" к галилеевскому, но, как ясно выразить это стремление, непонятно.

Здесь открывается важная педагогическая проблема: теория пределов и теория действительного числа, необходимые для строгого обоснования подобных операций, чересчур велики и серьезны для наших ограниченных возможностей. Кроме того, не следует надеяться, что современное изложение анализа избавляет читателя от необходимости понять, например, смысл выражения "x стремится к нулю". Мы выбираем вариант, в котором основная трудность понимания мгновенной скорости имеет наиболее отчетливый вид.

Итак, отношение пройденного за время o пути к продолжительности o равно

$$(2to + o^2)/o = 2t + o/2.$$

Поскольку o становится в конце концов нулем, то предельное отношение, т.е. мгновенная скорость, равна $2t$.

Критики справедливо указывали, что мы сначала считаем o не равным нулю и делим на o, а потом полагаем его равным нулю[4].

Тем не менее некорректные (это не значит приводящие к ошибкам) рассуждения дают полезные результаты. Мы здесь советуем понять ход мысли классиков и научиться с его помощью решать некоторый набор задач[5].

1.2. Производная

Почти определение 2. *Функцией будем называть любое правило, позволяющее по значению одной переменной (которую мы будем называть независимой переменной или аргументом функции) находить значение другой переменной (которую будем называть зависимой переменной или, допуская вольность, функцией). Чаще всего интересующие нас функции будут выражаться формулами, составленными из элементарных функций, к которым относятся степенные функции (в том числе и с показателями степени, заданными действительными числами), тригонометрические и показательные функции, а также логарифмы.*

Примеры функций: $y = x^{30}$, $y = x$, $y = sin(sin(sin(x)))$, $y = 2^x$. Здесь x независимая переменная, а y — зависимая. Имена переменных могут при необходимости меняться.

Когда мы пишем $y = f(x)$, мы подразумеваем, что вместо $f(x)$ может быть подставлена любая из формул, приведенных выше, и множество других.

Мы считаем, что правила графического изображения функций читателю известны. На рис. 1.1 изображены графики функций $y = x^2$ и $y = 2x - 1$.

---

[4] Особенно четкими были формулировки знаменитого философа архиепископа Дж. Беркли. Он демонстрировал, что рассуждения математиков, хотя и приводят к практически правильным результатам, но никак не могут считаться строгими. Целью его трактатов, разумеется, не было побудить математиков к строгому обоснованию анализа (хотя именно к этому привела его точная критика). Он хотел пресечь попытки математиков вторгаться в богословские вопросы. Аргументация его была проста — сначала научитесь правильно рассуждать в своей области, а затем уже учите рассуждать других.

[5] Отметим, что такие рассуждения не могут привести к формальному противоречию, но доказательство непротиворечивости лейбницева анализа оказывается не менее трудным, чем классическое обоснование современного анализа.

Упражнение 1.1. Проверить, что оба графика проходят через точку (1; 1) (абсцисса и ордината которой равны 1).

Упражнение 1.2. Доказать, что все точки параболы $y = x^2$ расположены выше точек прямой $y = 2x - 1$, имеющих ту же самую абсциссу.

Как сообщают историки, первые свои эксперименты над свободно падающими телами профессор университета города Пизы Г. Галилей проводил на знаменитой пизанской башне, которая и в те годы уже была наклонной, что создавало определенные удобства для исследования вертикального движения тел под действием силы тяжести.

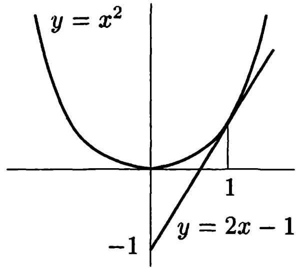


Рис. 1.1

Предположим, Галилей подбросил камень вертикально вверх со скоростью v м/сек. Построим график зависимости положения камня от времени (рис. 1.2). Примем за нуль ординату вершины башни, а за нулевой момент времени момент начала движения камня. Уравнение, связывающее высоту (выраженная в метрах зависимая переменная h) и время (выраженная в секундах независмая переменная t), таково:

$$y = vt - gt^2/2.$$

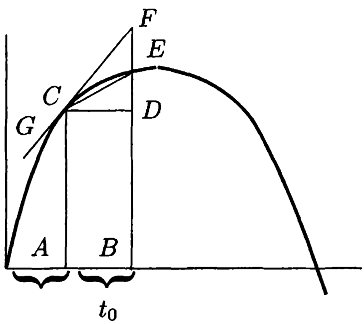


Рис. 1.2

Как и прежде, будем считать, что ускорение свободного падения на поверхности земли равно не 9,8 м/сек^2, а ровно 1 м/сек^2. Предположим также, что в начальный момент камню была придана скорость 2 м/сек. Тогда уравнение движения камня приобретет вид

$$y = 2t - t^2/2.$$

График этой функции представляет собой обращенную книзу параболу, проходящую через точку (0; 0) (см. рис. 1.2). По оси ординат откладывается высота камня над вершиной башни, по оси абсцисс — время полета.

Умножая фантастические предположения, вообразим, что, кроме приписываемого ему телескопа, Галилей изобрел еще некий антигравитон, который позволяет нейтрализовать земное тяготение в районе

башни. Через секунду после начала полета камня Галилей включает антигравитон, и камень перестает притягиваться к земле, продолжая полет только по инерции. Как будет выглядеть график его дальнейшего движения? В момент $t_0 = 1$, когда Галилей включил антигравитон, камень имел некоторую мгновенную скорость, которую мы умеем теперь вычислять, рассматривая соответствующее отношение при t неограниченно приближающемся к 0. Посмотрим, что соответствует этим операциям на нашем графике на рис. 1.2.

Точка A на графике соответствует моменту $t_0 = 1$, точка B — моменту t_0+t. Расстояние от вершины башни в эти моменты измеряется отрезками AC и $BE = BD + DE$, причем $BD = AC$.

За время, прошедшее от t_0 до $t_0 + t$, изображенное отрезком AB, камень пролетел расстояние DE. Отношение DE/AB представляло бы среднюю скорость камня за этот период времени (если бы Галилей не включил антигравитон, а камень двигался бы в поле земного тяготения). Что будет измерять это отношение, если t брать все меньше и меньше? Очевидно, среднюю скорость за все меньшие промежутки времени. В пределе мы вполне вправе расчитывать получить мгновенную скорость в момент $t_0 = 1$.

При приближении точки E к C угол между прямой CE и касательной GCF к графику функии будет уменьшатся и в пределе CE совпадет с касательной. При этом отношение DE/AB будет приближаться к DF/AB, а оно не зависит от величины AB, даже если будем брать последний отрезок все меньше и меньше — поскольку соответствующие треугольники всегда остаются подобными.

Отношение DF/AB представляет собой тангенс угла наклона касательной, проведенной в данной точке к графику функции.

Вспомним теперь, что Галилей на самом деле не позволил камню лететь по параболической траектории и включил антигравитон. Мгновенная скорость в данной точке сохраняется в последующие моменты, и график движения камня после момента, изображенного точкой C, представляет собой луч касательной CF.

Заметим, что здесь мы имеем касательную к графику, в отличие от разобранного выше случая камня, вращаемого на веревке и отпущенного в свободный полет, где касательная была реальной траекторией движения камня в пространстве.

В дальнейшем мы будем иметь дело именно с графиками функций, а не с реальными траекториями. Наша функция описывала движение точки вдоль прямой (вертикальной прямой, поскольку Галилей подбросил камень вверх). В этом случае последнее или предельное отношение

DE/AB при неограниченном приближении AB к нулю задает угол наклона графика, который бы имел место, если бы мгновенная скорость камня в данной точке сохранилась и камень двигался бы далее равномерно.

Если функция описывает какой-то иной процесс, то во всяком случае это предельное отношение можно считать скоростью роста функции в данной точке.

➽ **Почти определение 3.** *Это предельное отношение (или скорость роста функции в данной точке, или тангенс угла наклона касательной к графику функции в данной точке) называют производной функции в данной точке.*

Согласование падежей скорее требует сочетания "производная функция", чем "производная функции". Исторически первичным и было, по-видимому, первое сочетание. Оно означает следующее: если подсчитать в каждой точке области определения данной функции ее производную (тангенс угла наклона касательной в каждой точке), то получится новая функция — новое правило, позволяющее находить по значению аргумента значение зависимой переменной, равное этому тангенсу, в каждой точке своему.

На первый взгляд кажется, что найти тангенс угла наклона касательной в каждой из бесконечного числа точек — задача непосильная. Но для большинства интересующих нас функций это ощущение обманчиво. Например, как мы уже убедились, тангенс угла наклона касательной к графику функции $y = x^2$ в точке x_0 всегда равен $2x_0$ (чем больше x, тем пропорционально больше тангенс угла наклона касательной). Одним усилием мысли и, так сказать, одним росчерком пера мы можем получить всю производную функцию целиком, хотя вынуждены давать определение производной в отдельной точке. На рис. 1.3 изображены график функции $y = x^2$ и график ее производной $y = 2x$. Тангенс угла наклона касательной к параболе в точке, например, A равен ординате точки B.

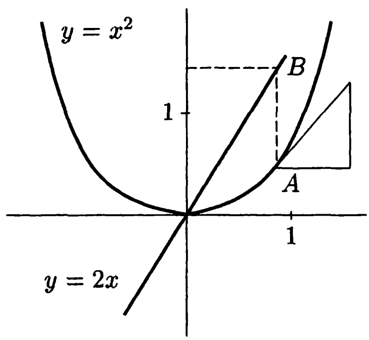


Рис. 1.3

Математический анализ столь поразительно эффективен именно потому, что одно рассуждение, реализующее одну какую-то идею, рабо-

тает в каждой точке из всей области определения функции, с которой имеет дело интересующая нас задача. Это и позволяет получать описания продолжительных непрерывных процессов, решать дифференциальные уравнения, описывать случайные величины с бесконечным множеством возможных значений и тому подобные вещи.

1.3. Производные от степенных функций

Проведем рассуждение в стиле Ньютона, чтобы найти производную от степенной функции $y = x^3$ в любой точке ее области определения. Рассмотрим значения функции в соседних точках x и $x + o$. Значение функции в точке $x + o$ преобразуем по формуле бинома того же Ньютона[6]

$$(x + o)^3 = x^3 + 3x^2o + 3xo^2 + o^3.$$

Тогда разность $(x + o)^3 - x^3$ примет вид

$$x^3 + 3x^2o + 3xo^2 + o^3 - x^3 = 3x^2o + 3xo^2 + o^3.$$

Разделим эту разность, как она представлена в последнем члене равенства, на расстояние между точками, в которых мы вычисляли значения функции, и получим среднюю скорость роста функции на этом промежутке (это расстояние равно $(x + o) - x$, то есть o).

$$(3x^2o + 3xo^2 + o^3)/o = 3x^2 + 3xo + o^2.$$

Только первый член в последнем выражении не содержит множитель o, поэтому в пределе при неограниченном приближении o к нулю от всего выражения останется только первый член $3x^2$, который характеризует мгновенную скорость роста функции $y = x^3$ в точке x. Наше рассуждение проводилось применительно к произвольной точке. Это значит, что тангенс угла наклона касательной к графику $y = x^3$ равен $3x^2$ при любом значении x.

Если хватит терпения, мы можем убедиться, что

— производная от функции $y = x^4$ равна $4x^3$,

— производная от функции $y = x^5$ равна $5x^4$,

и т.д., что можно записать общей формулой:

— производная от функции $y = x^n$ равна nx^{n-1}.

---

[6] Читатель, не знакомый с формулой бинома, может обратиться к главе 4 третьей части книги.

1.4. Производная функции $y = \sin x$, первый замечательный предел

Подобно разобранному выше примеру степенной функции, будем вычислять производную от функции $y = \sin x$, рассматривая отношение разности значений в двух соседних точках к расстоянию между этими точками:

$$(\sin(x + o) - \sin x)/o.$$

Мы можем преобразовать разность синусов по известному тригонометрическому тождеству[7]

$$\sin(x + o) - \sin x = 2\sin\frac{o}{2}\cos(x + \frac{o}{2}).$$

Тогда интересующее нас отношение можно переписать так:

$$\frac{\sin(x + o) - \sin x}{o} = 2\frac{\sin\frac{o}{2}\cos(x + \frac{o}{2})}{o} = \frac{\sin\frac{o}{2}}{\frac{o}{2}}\cos(x + \frac{o}{2}).$$

Как видим, наша задача несколько труднее, чем была в случае степенной функции. Нам нужно ответить на три вопроса.

Первый: что происходит с дробью $\frac{\sin o/2}{o/2}$ когда o неограниченно приближается к нулю?

Второй: что происходит с сомножителем $\cos(x + \frac{o}{2})$, когда o неограниченно приближается к нулю?

Третий: что происходит с произведением двух сомножителей, поведение каждого из которых при неограниченном приближении o к нулю известно?

1) Мы не будем приводить доказательства того факта, что $\frac{\sin o}{o}$ стремится к единице, когда o неограниченно приближается к нулю, а ограничимся убедительными рисунками. На рис. 1.4 рассматривается это отношение при различных значениях аргумента:

а) $o = 1$;

б) $o = 0{,}1$;

в) $o = 0{,}01$.

Напомним, что аргумент синуса мы выражаем в радианах. Для того чтобы вычислить синус аргумента o, надо отложить вдоль окружности единичного радиуса дугу AB, равную по длине o, от лежащей на окружности на одной горизонтали с ее центром точки A, а затем

---

[7] $\sin a - \sin b = 2sin\frac{a-b}{2}\cos\frac{a+b}{2}$.

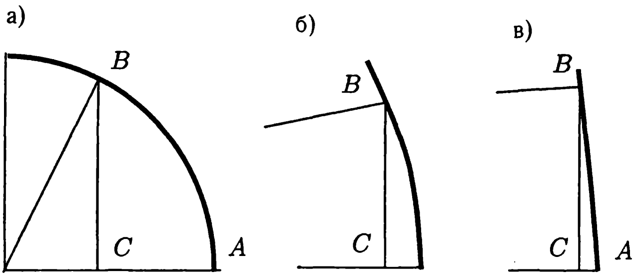


Рис. 1.4. Отношение дуги и соответствующего ей синуса при разных значениях аргумента: (а) $o = 1$; (б) $o = 0{,}1$ (увеличено в 10 раз); (в) $o = 0{,}01$ (увеличено в 100 раз).

найти ординату точки B, т.е. длину отрезка BC — это и есть синус дуги o (см. рис. 1.4). Очевидно, что разница длин дуги AB и отрезка BC становится все менее заметной при уменьшении o. Их отношение не зависит от масштаба рисунка и приближается к единице при неограниченном приближении аргумента o к нулю.

Мы будем говорить в таких случаях, что предел отношения $\frac{\sin o}{o}$, при o стремящемся к нулю, равен единице и записывать это формулой

$$\lim_{o\to 0} \frac{\sin o}{o} = 1.$$

Точный смысл понятия предела выражается следующими словами:

Литературное определение предела. *Говорят, что число A является пределом функции $f(x)$ при x стремящемся к x_0, если можно обеспечить как угодно малое отличие $f(x)$ от A, если только выбрать достаточно близкое к x_0 значение аргумента x.*

Мы будем в этом случае говорить также, что $f(x)$ стремится к A, когда x стремится к x_0. Можно заметить, что мы дали здесь несколько более общее определение, чем требуется для решения нашей задачи об отношении синуса к его аргументу — случай стремления o к нулю является частным по отношению к стремлению x к x_0.

Заметим также, что если мы доказали, что

$$\lim_{o\to 0} \frac{\sin o}{o} = 1,$$

то из этого можно заключить, что и

$$\lim_{o\to 0} \frac{\sin o/2}{o/2} = 1.$$

Аргументация при обосновании этого утверждения может опираться на литературное определение предела.

2) Следующий вопрос: что происходит с сомножителем $\cos(x + \frac{o}{2})$, когда o неограниченно приближается к нулю?

Почти определение непрерывной функции. *Глядя на график функции $y = \cos x$, изображенный на рис. 5, можно увидеть, что чем меньше разница между значениями аргумента, тем меньше различаются и значения функции. Этим свойством обладают все функции, график которых представляет собой непрерывную линию. Мы будем называть такие функции непрерывными. Точное определение непрерывности будет дано в главе 2.*

Если функция $f(x)$ непрерывна в точке x_0, то, сопоставляя данные выше "определения", можно заключить, что предел функции $f(x)$ при стремлении аргумента x к x_0 равен $f(x_0)$.

Таким образом и вторая задача решена: предел $\cos(x+\frac{o}{2})$ при неограниченном приближении o к нулю равен $\cos x$.

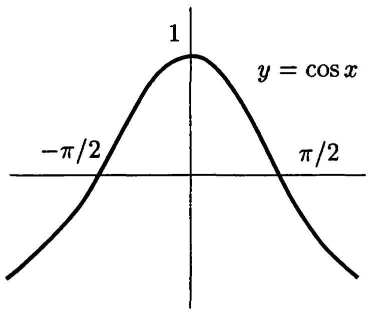


Рис. 1.5

Итак, в выражении

$$\frac{\sin \frac{o}{2}}{\frac{o}{2}} \cos(x + \frac{o}{2})$$

первый, дробный сомножитель стремится к единице, а второй к $\cos x$.

3) Можем ли мы заключить отсюда, что и все произведение стремится к $\cos x$? По-настоящему обоснованное заключение мы можем дать только уточнив определения и доказав соответствующие теоремы, чем мы займемся в следующей главе. Здесь же приведем лишь правдоподобные аргументы.

Поскольку предел при неограниченном приближении o к нулю

$$\frac{\sin \frac{o}{2}}{\frac{o}{2}}$$

равен единице (в том числе и для отрицательных значений аргумента o, поскольку и числитель и знаменатель меняют знак), то график этой

функции, который можно изобразить непрерывной линией везде, кроме точки $o = 0$, можно дополнить точкой (0; 1), т.е. положить функцию равной единице при $o = 0$.

Далее можно сослаться на то, что произведение непрерывных функций, конечно же, имеет непрерывный график.

В следующей главе мы докажем теоремы о пределах, из которых будет следовать, что сумма, разность и произведение непрерывных функций непрерывны и даже частное непрерывных функций непрерывно везде, где знаменатель не равен нулю.

Теперь вспомним, что все эти дополнительные сведения появились в ответ на вопрос, чему равна производная от функции $y = \sin x$. Мы посчитали предел отношения

$$(\sin(x + o) - \sin x)/o$$

и получили, что он равен $\cos x$. Это значит, что, проведя касательную к синусоиде в произвольной точке x ее области определения, мы можем быть уверены, что угловой коэффициент этой касательной в точности равен косинусу аргумента x.

Совершенно аналогично мы можем посчитать производную от функции $\cos x$, воспользовавшись тригонометрическим тождеством

$$\cos a - \cos b = -2 \sin \frac{a+b}{2} \sin \frac{a-b}{2}.$$

В этом случае

$$\frac{\cos(x+o) - \cos x}{o} = -2\frac{\sin \frac{o}{2} \sin(x + \frac{o}{2})}{o} = -\frac{\sin \frac{o}{2}}{\frac{o}{2}} \sin(x + \frac{o}{2})$$

при o стремящемся к нулю. Формула отличается от предыдущей знаком и вторым сомножителем. Результат: производная от функции $\cos x$ равна $-\sin x$.

1.5. Некоторые утверждения о производных

Функцию, имеющую производную в точке, будем называть дифференцируемой в данной точке. Функцию, дифференцируемую в каждой точке отрезка или интервала, будем называть дифференцируемой на отрезке или соответственно на интервале.

Мы можем условиться теперь о способе выражения: если $f(x)$ некоторая функция, то ее производную будем обозначать $f'(x)$, в таком

случае $f'(x_0)$ — это значение производной функции, взятое в точке x_0, оно же производная функции $f(x)$ в точке x_0.

Когда функция задана формулой, возможно иное обозначение. Например, уже выведенные нами формулы производной степенной функции и функции $\sin x$ могут быть записаны так:

$$(x^n)' = nx^{n-1},$$
$$(\sin x)' = \cos x.$$

Некоторые утверждения о производных легко доказываются в общем виде.

1) Производная от постоянной функции $f(x) = C$ есть тождественно равная нулю функция $y = 0$.

Рассмотрим отношение приращения функции к приращению аргумента. В данном случае это будет выражение

$$\frac{f(x+o) - f(x)}{o} = \frac{C - C}{o}.$$

Это выражение тождественно равно нулю, значит и предел его при стремлении o к 0 также равен нулю при любом значении x.

2) Если функцию умножить на константу, то и ее производная умножится на ту же константу.

Рассмотрим отношение приращения функции $Cf(x)$ к приращению аргумента:

$$\frac{Cf(x+o) - Cf(x)}{o} = C\,\frac{f(x+o) - f(x)}{o}.$$

Поскольку это отношение увеличилось в C раз, то также в C раз увеличится и предельное отношение при o стремящемся к 0.

Таким образом $(Cf(x))' = Cf'(x)$.

3) Если функция представляет собой сумму двух других функций, то и ее производная равна сумме соответствующих производных.

Разберем пример суммы $x^n + \sin x$.

Приращение суммарной функции считается по формуле

$$((x+o)^n + \sin(x+o)) - (x^n + \sin x).$$

Отношение приращений есть

$$\frac{(x+o)^n + \sin(x+o) - x^n - \sin x}{o} =$$

$$= \frac{(x+o)^n - x^n + \sin(x+o) - \sin x}{o} =$$

$$= \frac{(x+o)^n - x^n}{o} + \frac{\sin(x+o) - \sin x}{o}.$$

Предел верхнего отношения равен сумме пределов слагаемых в последней строке.

Таким образом, $(f(x) + g(x))' = f'(x) + g'(x)$.

4) Если функция представляет собой произведение двух других функций, то ее производная почти всегда не равна произведению производных.

Рассмотрим функцию $y = xx$. Поскольку $(x)' = 1$, произведение производных функций-сомножителей будет также равно единице. В то же время мы уже показали, что производная от функции $y = x^2$ равна $2x$, а не единице.

Рассмотрим приращение функции $y = xx$ в точке x_0, дав переменной x приращение o. В точке x_0 значение функции равно x_0x_0, в точке $x_0 + o$ значение равно $x_0x_0 + x_0o + ox_0 + oo$. Вычитая первое из второго, получаем, что приращение функции равно $x_0o + ox_0 + oo$. Не производя расчет производной этой функции, поскольку мы ее уже знаем, заметим, что приращение функции-произведения складывается из значения первого сомножителя, умноженного на приращение второго сомножителя, плюс произведение второго сомножителя, умноженного на приращение первого, плюс член oo, который оказывается пренебрежимо мал при переходе к пределу.

5) Мы можем догадаться теперь, что производная произведения $f(x)g(x)$ будет равна $f(x)g'(x) + f'(x)g(x)$. Доказательство — в следующей главе.

6) Производная от частного двух функций в точках, где функция в знаменателе не равна нулю, вычисляется по формуле

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}.$$

Доказательство также можно найти в следующей главе.

1.6. Производная и экстремум функции

Если функция достигает в некоторой точке максимума или минимума, касательная к ее графику будет горизонтальна, следовательно, про-

изводная функции в данной точке будет равна нулю. На этом свойстве производной основывается метод нахождения максимумов и миимумов.

Пример 3 (метод наименьших квадратов). Пусть даны три точки на числовой оси: a, b, c. Зададим зависящую от x функцию формулой $f(x) = (x-a)^2 + (x-b)^2 + (x-c)^2$. При каком x значение этой функции минимально?

Решение. Найдем производную $f'(x)$. Поскольку производная от функции $y = (x-a)^2 = x^2 - 2ax + a^2$ равна $2x - 2a$ (можно воспользоваться только что сформулированными утверждениями о производной суммы функций и производной функции, умноженной на костанту), то $f'(x) = 2x - 2a + 2x - 2b + 2x - 2c = 6x - 2(a+b+c)$. Производная функция существует на всей числовой оси, но лишь при $x_0 = (a+b+c)/3$ она равна нулю. Если функция $f(x)$ имеет минимум, то единственным претендентом на эту роль может быть только точка $(a+b+c)/3$, т.е. среднее арифметическое заданных чисел.

Наша функция $f(x)$ представляет собой сумму квадратов расстояний от точки x до точек a, b и c. Это значит, что функция не может не иметь минимума в какой-то точке числовой оси, следовательно, задача решена.

Метод наименьших квадратов будет использоваться в четвертой части книги, посвященной теории вероятностей и статистике.

Глава 2

Предел и производная

2.1. Техника ϵ и δ

В середине XIX века определение предела, по сути вполне эквивалентное данному в предыдущей главе, стали выражать на специальном языке. Его усвоение требует определенных усилий, которые мы не считаем обязательными для минимального знакомства с математическим анализом, поэтому данная тема вынесена в нашей книге в линию четных глав, ориентированных на более продвинутого читателя, готовящегося получить на экзамене отличную оценку.

Доказать теоремы о пределах, не используя этот язык, тоже можно, но такой "литературный" стиль ничуть не прояснит существо вопроса — наиболее адекватно теоремы о пределах выражаются на языке, окончательно утвердившемся в середине XIX века.

Напомним наше "литературное определение предела":

Говорят, что число A является пределом функции $f(x)$ при x стремящемся к x_0, если можно обеспечить как угодно малое отличие $f(x)$ от A, если только выбрать достаточно близкое к x_0 значение аргумента x.

Примерно с середины XIX века и до наших дней это определение выглядит так:

➽ **Определение 1′.** *Число A является пределом функции $f(x)$ при x стремящемся к x_0, если для любого сколь угодно малого положительного числа ϵ можно найти положительное число δ, такое, что неравенство $0 < |x - x_0| < \delta$ обеспечивает $|f(x) - A| < \epsilon$.*

Смысл последних неравенств следует понимать предельно просто:

— $0 < |x - x_0|$ означает, что x не должно равняться x_0. Например, в определении производной мы делим на $x - x_0$, поэтому обязаны исключить $x - x_0 = 0$;

— $|x - x_0| < \delta$ означает в точности, что расстояние между двумя значениями аргумента x и x_0 меньше δ;

— $|f(x) - A| < \epsilon$ означает, что числа $f(x)$ и A также отличаются мало, а именно меньше, чем на ϵ.

Таким образом, второе определение есть не что иное, как переформулировка первого. Выигрыш формулировки на языке δ и ϵ проявляется только при доказательстве теорем.

Специально для нынешнего компьютерного поколения читателей мы приведем модернизированное определение предела, которым и будем пользоваться далее:

➽ **Определение 1.** *Число A является пределом функции $f(x)$ при x стремящемся к x_0, если имеется алгоритм*[1]*, который для любого сколь угодно малого положительного числа ϵ, подаваемого ему на вход, выдает на печать число δ, такое, что при вычислении значения функции $f(x)$ для значений аргумента, лежащих от x_0 на расстоянии меньшем, чем δ (но не совпадающих с x_0), мы обязательно получим близкое к A значение $f(x)$, а именно отличающееся от A меньше, чем на ϵ.*

Покажем, как работает это определение.

Теорема 2.1. *Пусть при x стремящемся к x_0 предел $f(x)$ равен A, а предел $g(x)$ равен B. Тогда предел суммы $f(x) + g(x)$ равен $A + B$.*

РАЗЪЯСНЕНИЕ. Чтобы считать доказанным утверждение теоремы, мы должны предъявить алгоритм, который по заданному ϵ будет искать требуемое значение δ для функции $f(x) + g(x)$. Откуда он возьмется? Доказательством теоремы и будет построение этого алгоритма исходя из алгоритмов для $f(x)$ и $g(x)$, которые по условиям теоремы нам уже даны.

Доказательство. Наш алгоритм будет таков: пусть нам на вход подали ϵ.

---

[1] Под алгоритмом здесь можно понимать какой-то набор инструкций, выполняемых человеком с помощью компьютера или без такового, но так или иначе гарантирующих определенный числовой результат при наличии определенных входных числовых данных.

1) Запросим алгоритм для $f(x)$, подав ему на вход $\epsilon/2$. Этот алгоритм вернет нам число δ_1, гарантирующее, что $|f(x) - A| < \epsilon/2$ при $0 < |x - x_0| < \delta_1$.

2) Запросим алгоритм для $g(x)$, подав ему на вход $\epsilon/2$. Этот алгоритм вернет нам число δ_2, гарантирующее, что $|g(x) - B| < \epsilon/2$ при $0 < |x - x_0| < \delta_2$.

3) Напечатаем число δ, равное меньшему из чисел δ_1 и δ_2.

Теперь мы можем убедиться, что нужные условия выполнены. Если x отличается от x_0 меньше, чем на δ, то это значит, что их разница одновременно меньше и δ_1, и δ_2. Тогда $f(x)$ отличается от A меньше, чем на $\epsilon/2$ (это гарантирует первый алгоритм), а $g(x)$ отличается от B меньше, чем на $\epsilon/2$ (это гарантирует второй алгоритм).

Но тогда $f(x) + g(x)$ отличается от $A + B$ меньше, чем на ϵ. Действительно, во-первых,

$$(f(x) + g(x)) - (A + B) = f(x) - A + g(x) - B.$$

Далее воспользуемся тем, что абсолютная величина суммы двух чисел не больше, чем сумма их абсолютных величин: $|a + b| \leq |a| + |b|$. Тогда

$$|(f(x)+g(x))-(A+B)| = |(f(x)-A)+(g(x)-B)| \leq |f(x)-A|+|g(x)-B|.$$

Каждое из двух слагаемых в последнем выражении меньше половины ϵ, значит их сумма меньше ϵ.

Таким образом, мы сконструировали алгоритм, который требуется для того, чтобы объявить, что предел суммы $f(x) + g(x)$ равен $A + B$, если x стремится к x_0. **Теорема доказана.**

Упражнение 2.1. Небольшой вариацией приведенной в предыдущем доказательстве цепи неравенств доказать, что предел разности функций равен разности пределов этих функций.

Теорема 2.2. *Пусть при x стремящемся к x_0 предел $f(x)$ равен A, а C — некоторое постоянное число. Тогда предел произведения $Cf(x)$ равен CA.*

Доказательство. Получив на вход ϵ, запрашиваем алгоритм для $f(x)$, подав ему на вход $\epsilon/|C|$. Получив соответствующее δ, мы при $0 < |x-x_0| < \delta$ имеем $|f(x)-A| < \epsilon/C$, и поэтому $|Cf(x)-CA| = |C||f(x)-A| < \epsilon$.

Теорема доказана.

Прежде чем перейти к доказательству двух других теорем о пределах, докажем одну лемму. Будем пользоваться теперь сокращенной записью, введенной в конце предыдущей главы, в качестве замены выражения “предел при x стремящемся к x_0”.

Лемма 1$'$. *Пусть дана функция $f(x)$ и*

$$\lim_{x \to x_0} f(x) = A > 0.$$

Тогда в некоторой окрестности точки x_0 функция $f(x)$ ограничена сверху, т.е. для некоторой константы M и некоторого достаточно малого числа δ если $0 < |x - x_0| < \delta$, то $f(x) < M$.

В этой и следующей лемме имеется один важный аспект. В данном и в похожих на него случаях математический анализ исследует поведение функций вблизи определенной точки. Здесь важны только самые общие характеристики. В доказываемой нами лемме мы можем брать какую угодно маленькую окрестность и какую угодно большую константу: для доказательства утверждений о пределах, в которых будет использоваться данная лемма, это не важно.

Для доказательства леммы пошлем алгоритму, обеспечивающему условие леммы, запрос для $\epsilon = 1$ (можно было бы взять и 117). Алгоритм вернет нам некоторое число δ, такое, что $|f(x) - A| < 1$, если $0 < |x - x_0| < \delta$. Лемма фактически уже доказана: если $f(x)$ отличается от A меньше, чем на единицу, то $f(x)$ не может быть больше $A + 1$. Таким образом, если взять x, отличающийся от x_0 меньше, чем на полученное от алгоритма число δ (говорят “взять x, лежащий в δ-окрестности x_0”), то $f(x)$ будет меньше числа $M = A + 1$.

Очень небольшие вариации позволяют доказать лемму в таком виде.

Лемма 1. *Пусть дана функция $f(x)$ и*

$$\lim_{x \to x_0} f(x) = A.$$

Тогда в некоторой окрестности точки x_0 функция $f(x)$ ограничена сверху по абсолютной величине: $|f(x)| < M$.

Упражнение 2.2. Доказать лемму.

Теорема 2.3. *Пусть*

$$\lim_{x \to x_0} f(x) = A \text{ и } \lim_{x \to x_0} g(x) = B.$$

Тогда

$$\lim_{x \to x_0} f(x)g(x) = AB.$$

Доказательство. Как и в первой теореме, мы строим некоторое правило, как по данному нам ϵ искать требуемое значение δ, — на этот раз для функции $f(x)g(x)$.

1) Сначала используем лемму 1 и найдем число δ_1, гарантирующее, что $|g(x)| < M$ при $0 < |x - x_0| < \delta_1$ для некоторой константы M, которую мы можем взять достаточно большой, чтобы выполнялось неравенство $M > |A|$.

Далее, исходим из алгоритмов для $f(x)$ и $g(x)$, которые по условиям теоремы нам уже даны.

2) Запросим алгоритм для $f(x)$, подав ему на вход $\epsilon/2M$. Этот алгоритм вернет нам число δ_2, гарантирующее, что $|f(x) - A| < \epsilon/2M$ при $0 < |x - x_0| < \delta_2$.

3) Запросим алгоритм для $g(x)$, подав ему на вход $\epsilon/2M$. Этот алгоритм вернет нам число δ_3, гарантирующее, что $|g(x) - B| < \epsilon/2M$ при $0 < |x - x_0| < \delta_3$.

Напечатаем число δ, равное меньшему из чисел δ_1, δ_2 и δ_3.

Теперь мы можем убедиться, что нужные условия выполнены. Если x отличается от x_0 меньше, чем на δ, то это значит, что их разница одновременно меньше и δ_1, δ_2 и δ_3. Тогда $f(x)$ отличается от A меньше, чем на $\epsilon/2M$ (это гарантирует первый алгоритм), а $g(x)$ отличается от B меньше, чем на $\epsilon/2M$ (это гарантирует второй алгоритм). Одновременно $|f(x)| < M$.

Но тогда $f(x)g(x)$ отличается от AB меньше, чем на ϵ. Действительно, во-первых,

$$f(x)g(x) - AB = f(x)g(x) - Ag(x) + Ag(x) - AB.$$

Далее преобразуем

$$|f(x)g(x) - AB| = |(f(x)g(x) - Ag(x)) + (Ag(x) - AB)| \le$$

$$\le |f(x)g(x) - Ag(x)| + |Ag(x) - AB| =$$

$$= |(f(x) - A)g(x)| + |A(g(x) - B)| = |f(x) - A||g(x)| + |A||g(x) - B|.$$

Каждое из четырех выражений в последнем члене равенства заменяем на соответствующую оценку

$$|f(x) - A||g(x)| + |A||g(x) - B| < \frac{\epsilon}{2M}(M) + (|A|)\frac{\epsilon}{2M} = \epsilon/2 + \frac{|A|}{M}\epsilon/2 < \epsilon.$$

(Последнее неравенство выполнено благодаря выбору $M > |A|$ в п. 1 нашего доказательства.)

Таким образом, наш алгоритм действительно выдает по ϵ необходимое значение δ.

Теорема доказана.

Лемма 2. *Пусть*

$$\lim_{x \to x_0} f(x) = A \neq 0.$$

Тогда в некоторой окрестности точки x_0 функция $f(x)$ ограничена по абсолютной величине снизу: $|f(x)| > \frac{|A|}{2}$.

Доказательство. Запросим алгоритм для $f(x)$, подав ему на вход $|A|/2$. Для полученного δ выполнено: если $0 < |x - x_0| < \delta$, то $|f(x) - A| < |A|/2$, т.е. $f(x)$ ближе к A, чем к нулю, а это значит, что $|f(x)| > |A|/2$. **Лемма доказана.**

Теорема 2.4′. *Пусть*

$$\lim_{x \to x_0} f(x) = A \neq 0.$$

Тогда

$$\lim_{x \to x_0} \frac{1}{f(x)} = \frac{1}{A}.$$

Доказательство. Пусть нам на вход подается число ϵ.

Найдем по лемме 2 число δ_1, обеспечивающее $|f(x)| > |A|/2$, если $x \neq x_0$ и лежит в δ_1-окрестности x_0. Поделив единицу на обе части неравенства и поменяв знак неравенства, получим

$$\frac{1}{|f(x)|} < \frac{2}{|A|}.$$

Подадим на вход алгоритма для $f(x)$ число $\epsilon A^2/2$. Алгоритм вернет нам число δ_2, обеспечивающее $|A - f(x)| < \epsilon A^2/2$.

Тогда при $0 < |x - x_0| < \delta_1$ и $0 < |x - x_0| < \delta_2$ одновременно

$$\left|\frac{1}{f(x)} - \frac{1}{A}\right| = \frac{|A - f(x)|}{|f(x)A|} = |A - f(x)|\frac{1}{|f(x)|}\frac{1}{|A|} < \frac{\epsilon A^2}{2}\frac{2}{|A|}\frac{1}{|A|} = \epsilon.$$

Таким образом, если мы на запрос ϵ ответим меньшим из чисел δ_1, δ_2, то обеспечим этим

$$\left|\frac{1}{f(x)} - \frac{1}{A}\right| < \epsilon.$$

Алгоритм построен, **теорема доказана**.

Последняя теорема этой серии является простым следствием предыдущих.

Теорема 2.4. *Пусть*

$$\lim_{x\to x_0} f(x) = A \text{ и } \lim_{x\to x_0} f(x) = B \neq 0.$$

Тогда

$$\lim_{x\to x_0} \frac{f(x)}{g(x)} = \frac{A}{B}.$$

Доказательство.

$$\frac{f(x)}{g(x)} = f(x)\frac{1}{g(x)}.$$

Мы можем применить теперь теорему 2.4′ к $1/g(x)$ и теорему 2.2 к произведению $f(x)$ и $1/g(x)$.

Определение 2. *Функция $f(x)$ называется непрерывной в принадлежащей области определения этой функции точке x_0, если*

$$\lim_{x\to x_0} f(x) = f(x_0).$$

Хотя при изучении графиков элементарных функций это свойство непрерывности графика кажется тривиальным для всех элементарных функций, для доказательства всякий раз надо строить соответствующий алгоритм. Мы построим для примера алгоритм, преобразующий ϵ в δ для функции $y = \cos x$.

Поскольку $\sin x$ измеряется перпендикуляром к оси абсцисс, а x измеряется дугой, выходящей из той же точки и оканчивающейся на той же оси, т.е. криволинейной наклонной, то $|\sin x| \leq |x|$. Это позволяет оценить модуль разности

$$|\cos x - \cos x_0| =$$

$$= |-2| \left|\sin\frac{x-x_0}{2}\right| \left|\sin\frac{x+x_0}{2}\right| \leq 2\frac{|x-x_0|}{2}\left|\sin\frac{x+x_0}{2}\right| \leq |x-x_0|.$$

Поэтому если $|x - x_0| < \delta = \epsilon$, то $|\cos x - \cos x_0| < \epsilon$. Алгоритм тривиален: получив на вход ϵ вернуть $\delta = \epsilon$.

Замечание 1. Похожим образом можно доказать, что непрерывны функции x^α, $\operatorname{tg} x$, $\operatorname{ctg} x$, $\operatorname{arctg} x$, $\operatorname{arcctg} x$, $\arcsin x$, $\cos x$, $\arccos x$, а также a^x и $log_a x$ для любой константы a.

2.2. Производная

По сравнению с данным в предыдущей главе определением производной, нижеследующее отличается только использованием строгого определения предела.

Пусть дана функция $f(x)$, заданная в окрестности точки $x = 17$. Рассмотрим новую функцию

$$g(x) = \frac{f(x) - f(17)}{x - 17}.$$

Эта новая функция может быть без помех вычислена во всякой точке x, в которой задана функция $f(x)$, кроме точки $x = 17$, где знаменатель обращается в нуль и прямо воспользоваться правилом для $g(x)$ невозможно. Несмотря на это, в некоторых случаях существует предел

$$\lim_{x\to 17} g(x) = \lim_{x\to 17} \frac{f(x) - f(17)}{x - 17}.$$

Упражнение 2.3. Проверить, что в определении предела функции при x стремящемся к x_0 ничего не говорится о значении функции в точке x_0.

Рассмотрим пример: для функции $f(x) = x^2$ предстоит найти предел функции $g(x) = \frac{x^2-17^2}{x-17} = x + 17$, а это совсем не трудно: $\lim_{x\to 17} x + 17 = 34$. Алгоритм для этого предела, выдающий δ по заданному ϵ таков: $\delta = \epsilon$.

Кроме точки 17 подобные пределы можно рассматривать во всех точках, в окрестности которых задана функция $f(x) = x^2$, т.е. на всей числовой оси.

➽ **Определение 3**. *Производной функции $f(x)$ в точке x_0 называется число*

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Пример 1. В предыдущей главе мы убедились, что

$$\lim_{x\to 0} \frac{\sin x}{x} = 1. \tag{2.1}$$

Используя этот факт, мы можем вычислить производную от функции $y = \sin x$ в произвольной точке x_0.

$$\lim_{x\to x_0} \frac{\sin x - \sin x_0}{x - x_0} = \lim_{x\to x_0} \frac{2\sin\frac{x-x_0}{2}\cos\frac{x+x_0}{2}}{x - x_0}.$$

По замечанию 1

$$\lim_{x\to x_0}\cos\frac{x+x_0}{2}=\cos\frac{x_0+x_0}{2}=\cos x_0.$$

По формуле (2.1)

$$\lim_{x\to x_0}\frac{\sin\frac{x-x_0}{2}}{\frac{x-x_0}{2}}=1.$$

Применяя теорему 3 о пределе произведения, получаем

$$\lim_{x\to x_0}\frac{2\sin\frac{x-x_0}{2}\cos\frac{x+x_0}{2}}{x-x_0}=\lim_{x\to x_0}\frac{\sin\frac{x-x_0}{2}}{\frac{x-x_0}{2}}\lim_{x\to x_0}\cos\frac{x+x_0}{2}=\cos x_0.$$

Замечание 2. Утверждение

$$\lim_{x\to x_0}\frac{\sin\frac{x-x_0}{2}}{\frac{x-x_0}{2}}=1$$

не следует непосредственно из формулы (2.1).

Для ее вывода надо построить новый $\epsilon-\delta$-алгоритм, исходя из алгоритма для предела (2.1). В данном случае этот алгоритм предельно прост: полученное от старого алгоритма δ надо поделить на 2, что обеспечит нужное неравенство.

Теперь мы можем сказать, что функция $\cos x$ является производной для функции $\sin x$. Это означает, что, вычисляя согласно определению число

$$\lim_{x\to x_0}\frac{\sin x-\sin x_0}{x-x_0},$$

мы всегда будем получать $\cos x_0$.

2.3. Некоторые теоремы о производной

Далее все функции мы предполагаем дифференцируемыми.

Теорема 2.5. *Производная от постоянной функции $f(x)=C$ есть тождественно равная нулю функция $y=0$.*

$$\frac{f(x)-f(x_0)}{x-x_0}=\frac{C-C}{x-x_0}=0.$$

Это выражение тождественно равно нулю, значит, и предел его при стремлении x к x_0 также равен нулю при любом значении x_0.

Теорема 2.6. *Если функцию умножить на константу, то и ее производная умножится на ту же константу:* $(Cf(x))' = Cf'(x)$.

Доказательство. Рассмотрим соответствующий предел отношения приращения функции к приращению аргумента:

$$\lim_{x\to x_0} \frac{Cf(x) - Cf(x_0)}{x - x_0} = \lim_{x\to x_0} C\,\frac{f(x) - f(x_0)}{x - x_0} = C\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Последнее равенство следует из теоремы 2 о пределах.

Теорема 2.7. *Если функция представляет собой сумму двух других функций, то и ее производная равна сумме соответствующих производных:* $(f(x) + g(x))' = f'(x) + g'(x)$.

Доказательство.

$$\lim_{x\to x_0} \frac{f(x) + g(x) - (f(x_0) + g(x_0))}{x - x_0} = \lim_{x\to x_0} \frac{f(x) - f(x_0) + g(x) - g(x_0)}{x - x_0} =$$

$$= \lim_{x\to x_0} \left(\frac{f(x) - f(x_0)}{x - x_0} + \frac{g(x) - g(x_0)}{x - x_0}\right) =$$

$$= \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} + \lim_{x\to x_0} \frac{g(x) - g(x_0)}{x - x_0}.$$

Последнее равенство имеет место по теореме 1 о пределе суммы функций.

Теорема доказана.

Теорема 2.8. *Если функция представляет собой произведение двух других функций, то ее производная вычисляется по формуле:* $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$.

Доказательство. Заметим, во-первых, что

$$f(x)g(x) - f(x_0)g(x_0) = f(x)g(x) - f(x_0)g(x) + f(x_0)g(x) - f(x_0)g(x_0) =$$

$$= (f(x) - f(x_0))g(x) + f(x_0)(g(x) - g(x_0)).$$

Применим последнее равенство к отношению приращений:

$$\lim_{x\to x_0} \frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0} =$$

$$= \lim_{x\to x_0}\left(\frac{(f(x)-f(x_0))g(x)}{x-x_0}+\frac{f(x_0)(g(x)-g(x_0))}{x-x_0}\right).$$

Дальше применим теоремы о пределах суммы и произведения

$$\lim_{x\to x_0}\left(\frac{f(x)-f(x_0)}{x-x_0}g(x)+f(x_0)\frac{g(x)-g(x_0)}{x-x_0}\right)=$$

$$=\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}\lim_{x\to x_0}g(x)+f(x_0)\lim_{x\to x_0}\frac{g(x)-g(x_0)}{x-x_0}.$$

Заменяя выражения со знаком “lim” их значениями, получаем $f'(x_0)g(x_0)+f(x_0)g'(x_0)$.

Теорема доказана.

Теорема 2.9. *Производная от частного двух функций в точках, где функция в знаменателе не равна нулю, вычисляется по формуле*

$$\left(\frac{f(x)}{g(x)}\right)'=\frac{f'(x)g(x)-f(x)g'(x)}{(g(x))^2}.$$

Доказательство. Заметим, что

$$\frac{f(x)}{g(x)}-\frac{f(x_0)}{g(x_0)}=\frac{f(x)g(x_0)-f(x_0)g(x)+(-f(x_0)g(x_0)+f(x_0)g(x_0))}{g(x)g(x_0)}=$$

$$=\frac{f(x)g(x_0)-f(x_0)g(x_0)-f(x_0)g(x)+f(x_0)g(x_0)}{g(x)g(x_0)}=$$

$$=\frac{f(x)-f(x_0)}{g(x)}-f(x_0)\frac{g(x)-g(x_0)}{g(x)g(x_0)}.$$

Соединив первое и последнее выражения этой цепочки равенств и поделив их на $x-x_0$, перейдем к пределу при x стремящемся к x_0, чтобы получить искомое значение производной от частного функций $f(x)$ и $g(x)$.

$$\lim_{x\to x_0}\left(\frac{f(x)}{g(x)}-\frac{f(x_0)}{g(x_0)}\right)/(x-x_0)=$$

$$=\lim_{x\to x_0}\left(\frac{f(x)-f(x_0)}{g(x)}-f(x_0)\frac{g(x)-g(x_0)}{g(x)g(x_0)}\right)/(x-x_0)=$$

$$=\lim_{x\to x_0}\left(\frac{f(x)-f(x_0)}{g(x)(x-x_0)}-f(x_0)\frac{g(x)-g(x_0)}{g(x)g(x_0)(x-x_0)}\right)=$$

$$= \lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}\lim_{x\to x_0}\frac{1}{g(x)} - f(x_0)\lim_{x\to x_0}\frac{g(x)-g(x_0)}{x-x_0}\lim_{x\to x_0}\frac{1}{g(x)g(x_0)}.$$

Мы многократно применяли здесь доказанные теоремы о пределах.

Теперь заменим выражения, содержащие знак "lim", на соответствующие значения. Получим следующий эквивалент последнего члена равенства:

$$\frac{f'(x_0)}{g(x_0)} - f(x_0)\frac{g'(x_0)}{(g(x_0))^2}.$$

Осталось привести дроби к общему знаменателю:

$$\frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2}.$$

Это и есть искомая формула.

Теорема доказана.

2.4. Производная и экстремум функции

Предположим, что функция $f(x)$ имеет в точке x_0 минимум. Можем ли мы быть уверены, что касательная к графику $f(x)$ в этой точке горизонтальна, т.е. ее производная $f'(x_0) = 0$? Не может ли так случиться, что касательная наклонна, а функция "успевает" сделать крутой поворот, как это показано на рис. 2.1 а. Следующая теорема показывает, что этого быть не может и что при надлежащем увеличении масштаба график будет выглядеть так, как показано на рис. 2.1 б. Докажем предварительно лемму, очень похожую на лемму 2 предыдущего параграфа.

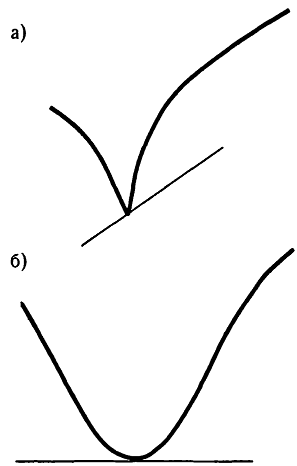


Рис. 2.1

Лемма 3. *Пусть*

$$\lim_{x\to x_0} f(x) = A > 0.$$

Тогда в некоторой окрестности точки x_0 *функция* $f(x)$ *положительна и даже ограничена снизу:* $f(x) > \frac{A}{2} > 0$.

Доказательство. Запросим алгоритм для $f(x)$, подав ему на вход $A/2$. Для полученного δ выполнено: если $0 < |x-x_0| < \delta$, то $|f(x)-A| < A/2$, т.е. $f(x)$ ближе к A, чем к нулю, а это значит, что $f(x) > A/2$. Лемма доказана.

Теорема 2.10. *Пусть $f'(x_0) = A > 0$. Тогда функция $f(x)$ возрастает в точке x_0, т.е. можно найти такое число δ, что $f(x) < f(x_0)$ левее точки x_0 и $f(x) > f(x_0)$ правее x_0, если только $|x - x_0| < \delta$.*

Доказательство. По определению производной

$$A = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Обозначим

$$g(x) = \frac{f(x) - f(x_0)}{x - x_0},$$

тогда

$$\lim_{x \to x_0} g(x) = A > 0.$$

По лемме 3 найдется число δ, такое, что при $0 < |x - x_0| < \delta$

$$g(x) = \frac{f(x) - f(x_0)}{x - x_0} > 0.$$

Это значит, что числитель и знаменатель дроби имеют одинаковые знаки, и если $x > x_0$, то $f(x) > f(x_0)$, а если $x < x_0$, то $f(x) < f(x_0)$.

Теорема доказана.

Совершенно аналогично доказывается утверждение об убывании функции $f(x)$ при $f'(x_0) = A < 0$.

Замечание 3. Глядя на график, можно предположить, что верно и более сильное утверждение: для любых двух точек из δ-окрестности x_0 функция $f(x)$ принимает меньшее значение в левой точке и большее в правой. Это не так. Опровергающий это утверждение пример: функция $f(x) = x^2 \sin(\frac{1}{x^2}) + x$ при $x \neq 0$, доопределенная значением $f(0) = 0$. Как угодно близко к точке $x = 0$ встречаются участки роста и убывания этой функции, хотя $f'(0) = 1$. Доказательство любознательный читатель легко проведет после овладения правилом дифференцирования сложной функции (см. раздел 6.1).

Такие примеры показывают, что математику не следует слишком доверять рассуждениям, опирающимся на очевидность графиков. В то же время видно, что пример не принадлежит к функциям, задаваемым одной формулой, которые в наибольшей степени согласуются с нашей интуицией функции. Спор между сторонниками и противниками допущения подобных функций в математический анализ был решен в пользу первых в XVIII веке. Психолог, к сожалению, тоже не может ограничиться рассмотрением только элементарных функций.

Из теоремы 10 следует, что если дифференцируемая функция $f(x)$ имеет максимум или минимум в точке x_0 и имеет в этой точке производную $f'(x_0)$, то $f'(x_0) = 0$. Если бы производная не равнялась нулю, то по теореме 10 функция бы возрастала или убывала в точке x_0, а значит, не могла бы иметь ни максимума, ни минимума.

Это рассуждение дает средство отыскания минимумов и максимумов функций. Однако условие $f'(x_0) = 0$ является необходимым, но не достаточным. Функция $y = x^3$ имеет нулевую производную в точке 0, но вообще не имеет минимумов и максимумов. Наиболее простым достаточным условием максимума является следующий: в рассматриваемой точке производная меняет знак. Если левее данной точки производная положительна (функция растет), а правее — отрицательна (функция убывает), то в данной точке имеется максимум. Минимум характеризуется противоположной сменой знака.

Если производная функция также дифференцируема, то достаточное условие можно переформулировать. Производная от производной функции называется второй производной (аналогично можно говорить о третьей, четвертой и т.д. производных). Если первая производная левее данной точки была положительной и, перейдя через 0, стала правее нее отрицательной, то значит она убывала в окрестности этой точки, т.е. ее производная (а это и есть вторая производная исходной функции) отрицательна. Таким образом, если первая производная функции в данной точке равна нулю, а вторая меньше нуля, то функция имеет в данной точке максимум, если первая равна нулю, а вторая больше нуля, то функция имеет минимум в данной точке.

Упражнение 2.4. Доказать, что функция $\cos x$ имеет максимум в точке 0.

В основе дальнейших рассуждений о связи поведения функции и ее производной лежит следующая теорема:

Теорема Ролля. *Пусть $f(x)$ дифференцируемая на отрезке $[a; b]$ функция, причем $f(a) = f(b)$. Тогда существует точка c, принадлежащая $[a; b]$, в которой производная обращается в нуль: $f'(c) = 0$.*

Идея доказательства. Функция $f(x)$ либо является константой, либо не является. В первом случае любая точка отрезка может быть выбрана в качестве c (теорема 2.5). Если же наша функция не постоянна, то из графических соображений у нее есть либо максимум, либо минимум. В качестве c и надо взять точку, где функция достигает максимума или минимума.

> **Замечание 4**. Для того чтобы приведенное выше рассуждение стало доказательством, надо доказать, что непрерывная на отрезке функция достигает минимального и максимального значений. Это оказывается довольно трудным делом и требует точной теории действительного числа. Читатель может найти ее изложение в любом учебнике математического анализа, адресованном математикам и техникам.

Теорема Лагранжа. *Пусть $f(x)$ дифференцируемая на отрезке $[a;b]$ функция. Тогда существует точка c, принадлежащая $[a;b]$, в которой*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Доказательство. Сведем теорему Лагранжа к теореме Ролля. Рассмотрим несколько "подправленную" линейным слагаемым функцию

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

При $x = a$ значения функций совпадают $g(a) = f(a)$, поскольку равно нулю выражение в последней скобке. При $x = b$ второе слагаемое компенсирует рост функции $f(x)$ и $g(b)$ снова равно $f(a)$. Тем самым $g(a) = g(b)$.

К функции $g(x)$ можно применить теорему Ролля, поэтому $g'(c) = 0$ для некоторой точки c. Найдем производную от функции $g(x)$:

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

Если $g'(c) = 0$, то

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Теорема доказана.

Упражнение 2.5. С помощью теорем о производной доказать, что

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

Следствие 1 из теоремы Лагранжа. *Если всюду на отрезке $[a; b]$ $f'(x) = 0$, то $f(x)$ постоянная функция.*

Доказательство. Если неверно, что $f(x) = C$ для всех x, то существуют две точки, в которых функция имеет разные значения: $f(a_1) \neq f(b_1)$. По теореме Лагранжа между ними найдется точка c, в которой

$$f'(c) = \frac{f(b_1) - f(a_1)}{b_1 - a_1}.$$

Поскольку $f(a_1) \neq f(b_1)$, то $f'(c) \neq 0$, что противоречит условию. Следовательно, точек с различными значениями на отрезке нет и функция постоянна.

Следствие доказано.

Следствие 2 из теоремы Лагранжа. *Если у функций $f(x)$ и $g(x)$ на всем отрезке $[a; b]$ совпадают производные, то они отличаются на константу: $f(x) - g(x) = C$.*

Доказательство. По теореме о производной разности функций, производная разности $f(x) - g(x)$ равна разности их производных, т.е. равна нулю всюду на отрезке. По только что доказанному следствию, разность этих функций постоянна.

Следствие доказано.

Глава 3

Определенный интеграл (идеи и примеры)

В первой главе мы сформулировали парадокс кругового движения:

Если продолжительность полета с данной мгновенной скоростью не равна нулю, то, двигаясь по касательной, тело покинет окружность. Если продолжительность полета с данной мгновенной скоростью равна нулю, то либо (1) тело никуда не сдвинется, либо (2) движение должно получаться суммированием бесконечного числа нулевых сдвигов.

Мы, вместе с математиками XVII века, остановились на ответе (2). Добавим теперь к этому, что нули бывают разные не только по направлению, но и по величине.

Парадокс II. *Если мгновенная скорость в некоторый момент велика, то ее "нулевой" вклад в бесконечную сумму будет больше "нулевого" вклада момента, в который мгновенная скорость меньше.*

Оба парадокса касаются синтеза целой траектории движения из мгновенных скоростей, данных в каждый момент времени. Что касается обратной задачи — отыскания мгновенных скоростей, если дана траектория, то после того, как дано определение производной, она выглядит более понятной.

Идея Ньютона и Лейбница состояла в том, что задачу синтеза можно решать как обратную к задаче отыскания производной. Действительно, пусть нам даны мгновенные скорости некоторого тела в каждый момент времени. Если найти такую траекторию, чтобы ее производ-

ная как раз совпадала с данными нам мгновенными скоростями, то, не решая задачу синтеза непосредственно, мы получим ответ.

Например, если мгновенные скорости прямолинейно движущегося тела зависят от времени по закону $v(t) = t^3$, то пройденный телом путь можно вычислить по формуле $S(t) = t^4/4$ — именно потому, что $S'(t) = v(t)$.

Замечательно, что этот же подход позволяет считать площади фигур, ограниченных кривыми линиями.

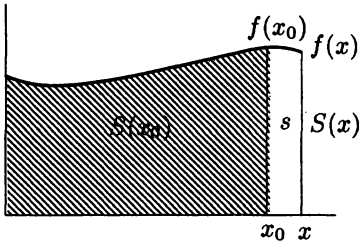


Рис. 3.1

На рис. 3.1 изображен график некоторой функции $y = f(x)$. Фигура, ограниченная сверху этим графиком, снизу осью абцисс, а по сторонам осью ординат и вертикальной прямой, проходящей через точку x на оси абсцисс, называется криволинейной трапецией. Обозначим ее площадь через $S(x)$. При разных значениях x площадь $S(x)$ имеет разные значения. Таким образом, $S(x)$ есть функция переменной x. Найдем производную от $S(x)$ в точке x_0.

$S(x_0)$ представляет собой площадь закрашенной фигуры, $S(x)$ площадь фигуры, полученной прибавлением маленькой криволинейной трапеции, помеченной малой буквой s. Составим отношение приращения площади к приращению аргумента.

$$\frac{S(x) - S(x_0)}{x - x_0} = \frac{s}{x - x_0}\,.$$

Что будет с фигурой s, если x бесконечно приближается к x_0? Поскольку при этом $f(x)$ все меньше отличается от $f(x_0)$, разброс ординат точек на верхней границе фигуры делается все меньше, так что в близком к предельному положении эта фигура становится почти что прямоугольником с основанием $x - x_0$ и высотой $f(x_0)$. Площадь этого прямоугольника $s = f(x_0)(x - x_0)$. Это значит, что

$$\frac{S(x) - S(x_0)}{x - x_0} = \frac{s}{x - x_0}$$

в пределе равно $\frac{f(x_0)(x-x_0)}{x-x_0}$, или $f(x_0)$, Таким образом, производная $S'(x_0) = f(x_0)$.

Это равенство верно в любой точке, поэтому $S'(x) = f(x)$. Теперь мы можем легко найти любую площадь, ограниченную графиком функций $f(x)$, горизонтальной прямой $y = 0$ и вертикальными прямыми $x = a$, $x = b$.

Для того чтобы вычислить площадь криволинейной трапеции, надо найти какую-нибудь функцию $F(x)$, производная которой равна $f(x)$. Искомая площадь будет равна $F(b) - F(a)$. Эта формула называется формулой Ньютона—Лейбница.

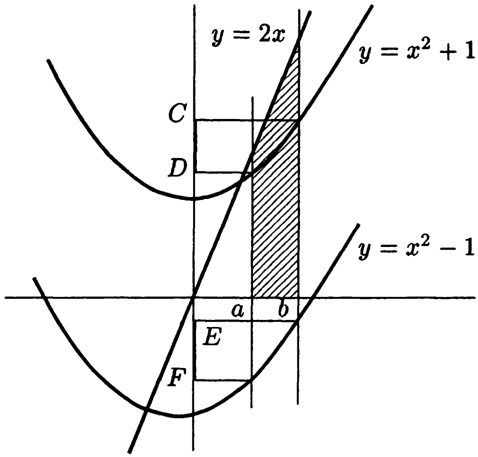


Рис. 3.2

Формула может показаться странной, поскольку существует бесконечно много функций, имеющих одинаковую производную, а в нашем рецепте не было сказано, какую из них выбирать. Например, обе функции $y = x^2 + 1$ и $y = x^2 - 1$ имеют одну и ту же производную $y = 2x$. По нашему рецепту получается, что обе они могут использоваться для вычисления площади трапеции, ограниченной осью абсцисс, горизонтальной прямой $y = 2x$ и вертикальными прямыми $x = a$ и $x = b$ (см. рис. 3.2).

Действительно, мы можем вычислить площадь трапеции прямо, пользуясь тем, что значение линейной функции $y = 2x$ в точках a и b (т.е. длины оснований уложенной "набок" трапеции) равны соответственно $2a$ и $2b$, а высота трапеции равна $b - a$. Таким образом, $S = (b - a)(2a + 2b)/2 = b^2 - a^2$.

По формуле Ньютона—Лейбница, используя функцию $y = x^2 + 1$, получаем $S = (b^2 + 1) - (a^2 + 1)$, а используя функцию $y = x^2 - 1$, получаем $S = (b^2 - 1) - (a^2 - 1)$.

Ответы совпадают. Объясняется это тем, что любые две функции, имеющие одинаковую производную, отличаются на константу (в разобранном нами случае эта константа равна двум). При вычитании значений на правой и левой границе области эти константы взаимно уничтожаются. На нашем чертеже разности ординат парабол в точках b и a

изображаются выделенными отрезками CD и EF, которые очевидно равны.

Нахождение производной данной функции называется дифференцированием. Обратная задача — нахождение функции $F(x)$, чья производная равна данной $f(x)$, — интегрированием. Такая функция $F(x)$ называется первообразной для $f(x)$, а площадь фигуры, ограниченной графиком функции $f(x)$, вертикальными прямыми $x = a$ и $x = b$ и осью абсцисс, определенным интегралом от a до b функции $f(x)$. Последнее выражение обозначается также формулой

$$\int_a^b f(x)\,dx.$$

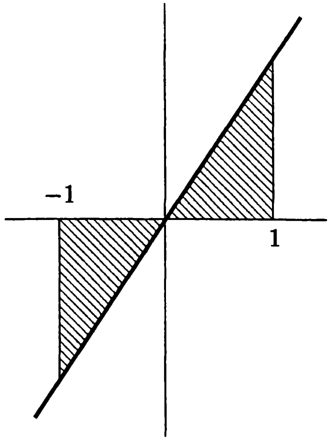


Рис. 3.3

a и b называются соответственно нижним и верхним пределами интегрирования. Пока знак dx в определенном интеграле будем считать только лишь указателем на то, что именно значения переменной x откладываются по оси абсцисс при построении фигуры, площадь которой нас интересует.

В новых обозначениях формула Ньютона—Лейбница приобретает вид:

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Замечание 1. Если $f(x)$ отрицательна, то площадь под осью абсцисс надо брать со знаком "минус". Например,

$$\int_{-1}^{1} 2x\,dx = 0,$$

поскольку график функции $y = 2x$ симметричен относительно начала координат (см. рис. 3.3), площади симметричных закрашенных треугольников равны, но левую площадь надо брать со знаком "минус".

Замечание 2. Хотя обычно предполагается, что нижний предел интегрирования расположен на оси абсцисс левее верхнего, имеет смысл принять удобное соглашение, которое позволяет выполнять некоторые операции правильно, не задумываясь:

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

Отметим, что и правая, и левая части равенства равны $F(b) - F(a)$.

Упражнение 3.1. Найти первообразную и, пользуясь формулой Ньютона—Лейбница, вычислить значение интеграла от 0 до -1 следующих функций:
1) $y = -1$;
2) $y = 2x$;
3) $y = -2x$.

Вычислив непосредственно площади фигур и использовав замечания 1 и 2, убедиться, что результаты, полученные двумя разными путями, совпадают.

Пример 1. В теории вероятностей и математической статистике формула Ньютона—Лейбница играет важную роль. Рассмотрим функцию

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

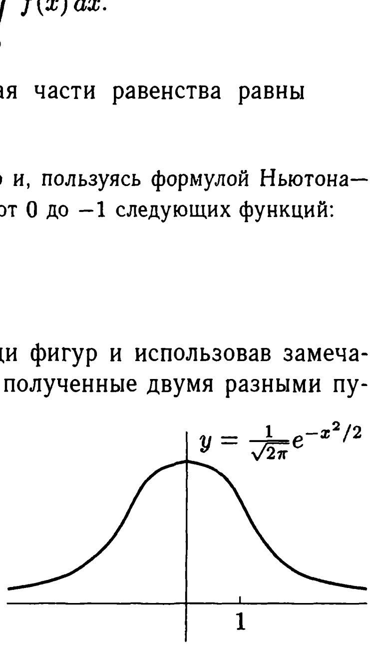


Рис. 3.4

(ее график изображен на рис. 3.4). С ее помощью описываются вероятности случайных событий. Для того чтобы оценить, например, вероятность того, что первая женщина, вошедшая на факультет психологии московского университета завтра после 12 часов дня, будет иметь рост между 170 и 180 см, надо посчитать определенный интеграл от 170 до 180 некоторого видоизменения функции $p(x)$.

Определим функцию $\Phi(x)$ следующим равенством:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-x^2/2}\,dx,$$

или короче:

$$\Phi(x) = \int_0^x p(x)\,dx.$$

Используя замечание 2, значение функции можно задать и при отрицательных значениях аргумента. Поскольку график функции $p(x)$ симметричен относительно оси ординат (функция четная: $p(x) = p(-x)$), то перемена знака при отрицательных значениях x приведет к тому, что $\Phi(x)$ будет нечетной: $\Phi(-x) = -\Phi(x)$.

По доказанному нами выше $\Phi'(x) = p(x)$, по крайней мере для положительных значений аргумента. Используя нечетность $\Phi(x)$ и четность $p(x)$, можно убедиться, что и при отрицательных значениях аргумента это равенство справедливо. Не слишком трудно, впрочем, доказать, что замечаний 1 и 2 достаточно, чтобы для

$$S(x) = \int_0^x f(x)\,dx$$

всегда выполнялось $S'(x) = f(x)$ для любых значений аргумента и любой непрерывной функции $f(x)$, независимо от их знака.

Функция $\Phi(x)$ обладает интересными свойствами:

1) $\Phi(x)$ монотонно растет. Если $x_1 < x_2$, то $\Phi(x_1) < \Phi(x_2)$.

2) При неограниченном росте x значение $\Phi(x)$ неограниченно приближается к $1/2$, при неограниченном убывании x значение $\Phi(x)$ неограниченно приближается к $-1/2$. Эти соотношения можно записать формулами

$$\lim_{x\to\infty} \Phi(x) = 1/2 \text{ и } \lim_{x\to-\infty} \Phi(x) = -1/2.$$

Благодаря этим формулам мы можем определить так называемые несобственные интегралы от $p(x)$:

$$\int_0^\infty p(x)\,dx = 1/2 \text{ и } \int_\infty^0 p(x)\,dx = 1/2.$$

Поскольку и формула Ньютона—Лейбница, и соображения сложения площадей говорят нам, что для любой функции $f(x)$ и любых пределов интегрирования

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx,$$

мы можем вычислить значение следующего интеграла:

$$\int_{-\infty}^{a} p(x)\,dx = \int_{-\infty}^{0} p(x)\,dx + \int_{0}^{a} p(x)\,dx = 1/2 + F(a).$$

Следующие упражнения 3.2 и 3.3 принадлежат к разряду совершенно обязательных для тех, кто собирается изучать математическую статистику.

Упражнение 3.2. Дать аналогичное определение интегралу

$$\int_{a}^{\infty} p(x)\,dx$$

и убедиться, что оно безошибочно работает как для положительных, так и для отрицательных значений a.

Функция $\Phi(x)$ не выражается через элементарные функции, поэтому вычисление ее значений представляет определенные трудности. В силу ее большой важности для статистических приложений ее значения приводятся в достаточно подробных таблицах. Для экономии места печатаются значения функции $\Phi(x)$ только для положительных значений аргумента.

Упражнение 3.3. Выразить через значения $\Phi(x)$ для положительных x следующие интегралы:

$$1)\ \int_{-\infty}^{-1} p(x)\,dx,$$

$$2)\ \int_{-1}^{1} p(x)\,dx,$$

$$3)\ \int_{-1}^{\infty} p(x)\,dx.$$

Глава 4

Определенный интеграл (доказательства)

Мы называем фигуру, ограниченную графиком непрерывной функции $f(x)$, осью абсцисс и вертикальными прямыми $x = a$ и $x = b$, криволинейной трапецией, образованной функцией $f(x)$ между точками a и b.

Мы не будем давать строгого определения площади такой фигуры, ограничившись интуитивным пониманием.

Нам понадобятся два очевидных свойства площадей:

1) Если $a = b$, то площадь соответствующей трапеции равна нулю.

2) Если фигура S_1 целиком лежит внутри фигуры S_2, то площадь первой фигуры меньше площади второй.

Мы будем рассматривать случай положительной функции. Случай функции, меняющей знак, рассматривается в основном аналогично, хотя и требует большей аккуратности.

Теорема 4.1. *Если*

$$\lim_{x\to x_0} f(x) = \lim_{x\to x_0} g(x) = A,$$

при этом $f(x) \leq g(x)$, а функция $h(x)$ всегда лежит между $f(x)$ и $g(x)$, т.е. $f(x) \leq h(x) \leq g(x)$, то

$$\lim_{x\to x_0} h(x) = A.$$

Доказательство. Запросим алгоритмы для данных двух пределов, подав им на вход ϵ. Обозначим δ меньшее из полученных от них чисел.

Тогда $|f(x)-A| < \epsilon$ и $|g(x)-A| < \epsilon$, если $0 < |x-x_0| < \delta$. Пусть x принадлежит δ-окрестности x_0, тогда $f(x)$ и $g(x)$ отличаются от A меньше, чем на ϵ, т.е. $A+\epsilon \leq f(x), g(x) \leq A-\epsilon$.

Это значит, что $A+\epsilon \leq f(x) \leq h(x) \leq g(x) \leq A-\epsilon$, откуда $h(x)$ также лежит между $A-\epsilon$ и $A+\epsilon$, т.е. $|h(x)-A| < \epsilon$. Алгоритм для $h(x)$ построен.

Теорема доказана[1].

Теорема 4.2. *Пусть $S(x)$ площадь криволинейной трапеции, образованной непрерывной положительной функцией $f(x)$ между точками a и x.*

Тогда $S(x)$, как функция переменной x, обладает свойством $S'(x) = f(x)$.

Доказательство. Доказательство будем вести на том же чертеже, что и эскиз доказательства в предыдущей главе (рис. 3.1). Пусть x_0 некоторая точка, $S(x_0)$ площадь соответствующей трапеции между a и x_0 (закрашенная фигура), x некоторая близкая к x_0 точка, а $S(x)$ площадь соответствующей трапеции, $s(x) = S(x) - S(x_0)$ — приращение площади.

На нашем чертеже функция $f(x)$ монотонна, но в общем случае она может много раз менять направление роста (и даже бесконечное число раз), поэтому доказательство приходится усложнять.

Рассмотрим функцию $f_{min}(x)$, равную минимуму $f(x)$ на отрезке $[x_0; x]$. Функция $f_{min}(x)$ имеет тот же самый предел, что и $f(x)$. Действительно, пусть на запрос ϵ алгоритм, обеспечивающий утверждение

$$\lim_{x\to x_0} f(x) = f(x_0),$$

отвечает числом δ. Тогда если для всех x, для которых $0 < |x-x_0| < \delta$, выполнено $|f(x) - f(x_0)| < \epsilon$, то это неравенство верно и для той точки на отрезке $[x_0; x]$, где $f(x)$ достигает минимума[2]. Это значит, что при том же δ верно и $|f_{min}(x) - f(x_0)| < \epsilon$.

[1] В математическом фольклоре эта теорема называется теоремой о двух милиционерах: если два милиционера, держа преступника под руки, идут в отделение милиции, то и он неминуемо туда попадет.

[2] В полном курсе анализа утверждения о том, что минимум и максимум непрерывной на отрезке функции достигаются в каких-то точках отрезка, обосновываются рядом теорем о действительных числах. Мы могли бы, впрочем несколько усложнив наше доказательство, вполне корректно обойтись и без этого утверждения о минимуме и максимуме.

Аналогично и предел функции $f_{max}(x)$, равной максимуму $f(x)$ на отрезке $[x_0; x]$, равен тому же значению $f(x_0)$.

При каждом x маленькая трапеция, помеченная буквой s, целиком лежит в прямоугольнике с тем же самым основанием $[x_0; x]$ и высотой $f_{max}(x)$ и целиком содержит прямоугольник с тем же основанием и высотой $f_{min}(x)$.

Для площадей это значит, что

$$(x - x_0)f_{min}(x) \le s(x) \le (x - x_0)f_{max}(x).$$

Поделив на $x - x_0$, получим

$$f_{min}(x) \le s(x)/(x - x_0) \le f_{max}(x).$$

Левый и правый члены неравенства стремятся к одному пределу, тем самым для функций $f_{min}(x)$, $s(x)/(x-x_0)$ и $f_{max}(x)$ выполнено условие теоремы 1, откуда

$$\lim_{x \to x_0} s(x)/(x - x_0) = f(x_0).$$

По определению производной $S'(x_0) = f(x_0)$.

Теорема доказана.

➽ **Определение 4.1.** *Первообразной для функции $f(x)$ называется любая функция $F(x)$, производная которой равна $f(x)$.*

Пусть нам надо вычислить площадь криволинейной трапеции, образованной функцией $f(x)$ между точками a и b. Как мы установили во второй главе, две функции, имеющие одну и ту же производную (теперь мы можем сказать: две первообразные функции $f(x)$), отличаются на постоянное число. Если мы нашли какую-то первообразную для $f(x)$, то можем быть уверены, что она отличается от искомой площади на константу, которая пока нам неизвестна.

Однако найти ее очень легко: $S(a) = 0$ (площадь фигуры, имеющей нулевую ширину), а это значит, что разность между $F(x)$ и $S(x)$ в точке a равна $F(a)$. Но эта разность постоянна при всех x, поэтому $S(x) = F(x) - F(a)$. В частности, $S(b)$, которая представляет площадь криволинейной трапеции между a и b, равна $F(b) - F(a)$.

Мы доказали формулу Ньютона—Лейбница.

Глава 5

Производные и неопределенные интегралы

5.1. Производные и неопределенные интегралы от элементарных функций

Как справедливо рассуждали творцы анализа, задача восстановления траектории по мгновенным скоростям и задача вычисления площади криволинейной трапеции под заданной функцией, если их решать с помощью предельных переходов, требует весьма изощренных усилий. Если же воспользоваться формулой Ньютона—Лейбница, то в огромном числе случаев задача решается без особых затрат, поскольку для многих функций мы знаем первообразные.

В таком случае понятно, что чем больше мы узнаем производных от разных функций, тем больше задач интегрирования мы сможем легко решать.

В приведенной ниже таблице собраны производные от элементарных функций. Сами функции занимают левый столбец, в среднем столбце помещены их производные.

В правом столбце расположены так называемые неопределенные интегралы, которые можно считать с помощью помещенных в той же

строке формул. Неопределенный интеграл это формула вида

$$\int f(x)\,dx.$$

Найти, или вычислить, неопределенный интеграл — это значит указать все множество первообразных для $f(x)$. Если нам удается найти одну из них, например, такую функцию $F(x)$, производная которой равна $f(x)$, то мы можем записать ответ:

$$\int f(x)\,dx = F(x) + C.$$

Константа C в правой части указывает на то, что, меняя ее, мы можем получить любую другую первообразную.

Например, можно записать

$$\int x^{15}\,dx = \frac{x^{16}}{16} + C.$$

Для проверки подобных равенств надо убедиться, что производная от правой части равна подынтегральной функции, что в нашем случае верно.

Упражнение 5.1. В строках 11, 12 и 13, 14 (см. табл. на стр. 176) помещены одинаковые интегралы с различными правыми частями. Чем объясняется эта неоднозначность вычисления

$$\int \frac{dx}{\sqrt{1-x^2}}?$$

Некоторые свойства неопределенного интеграла похожи на свойства производной и доказваются с их помощью. Взяв производные от обеих частей равенства, мы можем убедиться, что

$$\int Cf(x)\,dx = C\int f(x)\,dx \quad \text{и} \quad \int (f(x)+g(x))\,dx = \int f(x)\,dx + \int g(x)\,dx.$$

Интеграл от произведения функций не равен произведению интегралов, поскольку производная произведения не есть произведение производных.

Таблица производных от элементарных функций и интегралов, приводящих к элементарным функциям

1	$y = C$	$y' = 0$	
2	$y = x$	$y' = 1$	$\int dx = x + C$
3	$y = x^n$	$y' = nx^{n-1}$	$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C$
4	$y = \sin x$	$y' = \cos x$	$\int \cos x\,dx = \sin x + C$
5	$y = \cos x$	$y' = -\sin x$	$\int \sin x\,dx = -\cos x + C$
6	$y = a^x$	$y' = a^x \ln a$	$\int a^x\,dx = \frac{1}{\ln a}a^x + C$
7	$y = e^x$	$y' = e^x$	$\int e^x\,dx = e^x + C$
8	$y = \ln x$	$y' = \frac{1}{x}$	$\int \frac{1}{x}\,dx = \ln x + C$
9	$y = \operatorname{tg} x$	$y' = \frac{1}{\cos^2 x}$	$\int \frac{dx}{\cos^2 x} = \operatorname{tg} x + C$
10	$y = \operatorname{ctg} x$	$y' = -\frac{1}{\sin^2 x}$	$\int \frac{dx}{\sin^2 x} = -\operatorname{ctg} x + C$
11	$y = \arcsin x$	$y' = \frac{1}{\sqrt{1-x^2}}$	$\int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C$
12	$y = \arccos x$	$y' = -\frac{1}{\sqrt{1-x^2}}$	$\int \frac{dx}{\sqrt{1-x^2}} = -\arccos x + C$
13	$y = \operatorname{arctg} x$	$y' = \frac{1}{1+x^2}$	$\int \frac{dx}{1+x^2} = \operatorname{arctg} x + C$
14	$y = \operatorname{arcctg} x$	$y' = -\frac{1}{1+x^2}$	$\int \frac{dx}{1+x^2} = -\operatorname{arcctg} x + C$

Формулы в третьей строке верны не только для целых n, но и для любых действительных показателей степени.

5.2. Дифференцирование сложной функции и замена переменной в неопределенном интеграле

Кроме элементарных функций достаточно просто вычисляются производные так называемых сложных функций. Сложная функция получается композицией простых. Например, одна простая функция $y = \sin t$, другая $t = x^2$. Если подставить в первую вместо t его выражение через x, получим функцию $y = \sin(x^2)$. Как вычислять ее производную?

Сначала дадим правило.

1) Вычислить производную по t функции $y = \sin t$
($y'_t = \cos t$).

2) Вычислить производную по x функции $t = x^2$
($t'_x = 2x$).

3) Умножить первую на вторую
($y'_t t'_x = (\cos t)(2x)$).

4) Подставить в полученную формулу вместо t его выражение через x ($y'_t t'_x = (\cos x^2)(2x)$). Это и есть искомый результат: $y'_x = 2x\cos x^2$.

Это правило дает возможность находить производную от любой функции, заданной одной формулой. Дифференцирование даже такой функции:

$$y = \sqrt{\sin^{80} x + \arcsin^{-79} x}$$

не потребует ничего, кроме внимательной реализации шагов, предписанных приведенным выше правилом.

Пример 1. Найти производную функции $y = e^{-x^2}$.

Положим $y = e^t$; $t = -x^2$. Найдем производные: $y'_t = e^t$ и $t'_x = -2x$. Окончательно

$$y'(x) = e^t(-2x) = -2xe^{-x^2}.$$

Пытаясь совершить обратную операцию — найти первообразную для данной функции, — мы оказываемся в совершенно иной ситуации. Нет никакого единого метода нахождения первообразных даже для не слишком сложно устроенных функций. Мы дадим сейчас один прием сведения неизвестных интегралов к известным, который называется заменой переменных. Он позволяет решить большое количество задач на интегрирование.

Если нам надо найти интеграл

$$\int 2x\cos x^2\,dx,$$

то мы можем вспомнить, что десятью строками раньше мы нашли $(\sin x^2)' = 2x\cos x^2$. Это значит, что

$$\int 2x\cos x^2\,dx = \sin x^2 + C$$

— продифференцировав правую часть, получим подынтегральную функцию.

Оказывается, что довольно много задач на интегрирование можно решить, подобрав соответствующую сложную функцию, производная от которой равна подынтегральному выражению. Метод замены переменной в интеграле позволяет производить поиск сложной функции, имеющей нужную нам производную.

"Механизм" замены переменой опирается на одно из трудных понятий математического анализа — понятие дифференциала.

Почти определение дифференциала. Пусть $y = f(x)$ некоторая функция. Дифференциал этой функции в некоторой точке x обозначается dy и представляет собой выражение $f'(x)dx$. Таким образом, $dy = f'(x)dx$. Это соотношение позволяет делать замены переменных в неопределенных интегралах совершенно автоматически и тем не менее получать правильный результат.

Пусть нам надо вычислить интеграл

$$\int 2x\cos(x^2+1)\,dx.$$

Введем новую переменную $u = x^2+1$. Вычислим сначала производную этой функции: $(x^2+1)' = 2x$. Таким образом, $du = 2xdx$, и, поделив равенство на $2x$, получаем $dx = du/2x$. Подставим в интеграл выражение $du/2x$ вместо dx, а u вместо x^2+1. Получим

$$\int 2x\cos u\,du/2x = \int \cos u\,du.$$

Последний интеграл можно найти в таблицах

$$\int \cos u\,du = \sin u + C.$$

Теперь можно подставить x^2+1 вместо u и получить окончательный ответ

$$\int 2x\cos(x^2+1)\,dx = \sin(x^2+1) + C.$$

Найти правильную замену переменной совсем не просто. Нет никакого правила, указывающего, какую замену надо произвести. От читателя нечетных глав нашего учебника и не требуется преуспеть в этом искусстве, однако понять, как работает замена в тех нескольких случаях, которые разбираются в части 3 нашей книги, вполне реальная задача.

Наиболее простые замены из тех, что будут нам встречаться, представляют собой линейную функцию $u = ax + b$:

Пример 2. Найти интеграл

$$\int e^{3x+17}dx.$$

Делаем замену переменной $u = 3x + 17$, откуда $du = (3x + 17)'dx = 3dx$ или $dx = du/3$. Подставляя полученные выражения, получаем табличный интеграл

$$\frac{1}{3}\int e^u du = \frac{e^u}{3} = \frac{e^{3x+17}}{3}.$$

Пример 3. Найти интеграл

$$\int (x + 17)e^{(x+17)^2}dx.$$

Делаем замену переменной $u = (x + 17)^2$, откуда $du = 2(x + 17)dx$. После подстановки

$$\int \frac{1}{2}e^u du = \frac{e^{(x+17)^2}}{2} + C.$$

Глава 6

Производные от некоторых функций

6.1. Производная от сложной функции

Пусть y зависит от t и эта зависимость выражается функцией $y = f(t)$. Пусть t, в свою очередь, выражается через x функцией $t = g(x)$. Тогда композиция функций задает зависимость y от x: $y = f(g(x))$. Обозначим эту композицию $h(x) = f(g(x))$.

Производную этой функции вычисляют по формуле $h'(x) = f'_t(g(x))g'(x)$. Значок t снизу от буквы f указывает, что в данном случае надо "забыть" о том, что t является промежуточной переменной и дифференцировать $f(t)$ так, как мы бы дифференцировали функцию в уже привычном случае независимой переменной t.

Теорема 6.1. *Пусть функция $t = g(x)$ имеет производную в точке x_0, а функция $y = f(t)$ имеет производную в точке $g(x_0)$. Тогда функция $h(x) = f(g(x))$ имеет производную в точке x_0 и эта производная равна $f'_t(g(x_0))g'_x(x_0)$.*

Доказательство. Доказательство проведем для случая $g'(x_0) \neq 0$. По теореме 10 второй главы это значит, что функция либо возрастает, либо убывает в точке x_0, но во всяком случае $g(x) \neq g(x_0)$ в некоторой окрестности x_0.

Нам надо вычислить

$$\lim_{x \to x_0} \frac{h(x) - h(x_0)}{x - x_0}.$$

Преобразуем отношение приращений, используя $g(x) \neq g(x_0)$:

$$\frac{h(x) - h(x_0)}{x - x_0} = \frac{f(g(x)) - f(g(x_0))}{x - x_0} = \frac{f(g(x)) - f(g(x_0))}{g(x) - g(x_0)} \frac{g(x) - g(x_0)}{x - x_0}.$$

Мы имеем теперь произведение двух отношений, рассмотрим их пределы по отдельности. Если обозначить $t = g(x)$ и $t_0 = g(x_0)$, то

$$\lim_{x \to x_0} \frac{f(g(x)) - f(g(x_0))}{g(x) - g(x_0)} = \lim_{x \to x_0} \frac{f(t) - f(t_0)}{t - t_0}.$$

Если x стремится к x_0, то $t = g(x)$ стремится к $t_0 = g(x_0)$ В таком случае под знаком предела можно сделать замену и получить отношение

$$\lim_{t \to t_0} \frac{f(t) - f(t_0)}{t - t_0},$$

которое соответствует определению производной функции $f(t)$ в точке t_0 (это интуитивно ясное утверждение о дифференцируемой, а значит, непрерывной функции $g(x)$ можно доказать на языке $\epsilon - \delta$). Таким образом,

$$\lim_{x \to x_0} \frac{f(g(x)) - f(g(x_0))}{g(x) - g(x_0)} = f_t'(t_0) = f_t'(g(x_0)).$$

Предел второго отношения непосредственно соответствует определению производной $g(x)$ в точке x_0:

$$\lim_{x \to x_0} \frac{g(x) - g(x_0)}{x - x_0} = g'(x_0).$$

По теореме о пределе произведения искомый предел равен произведению пределов, которые дают $f_t'(g(x_0))$ и $g'(x_0)$, соответственно. То есть

$$h'(x) = \lim_{x \to x_0} \frac{f(g(x)) - f(g(x_0))}{x - x_0} = f_t'(g(x_0))g'(x_0).$$

Теорема доказана.

Доказательство в случае $g'(x_0) = 0$ требует некоторых дополнительных усилий. Мы ограничимся изложенным выше.

Пример 1. Найти производную от функции

$$y = \frac{1}{\sqrt{1 + x^2}}.$$

Разложим данную функцию в композицию двух функций: $f(t) = 1/\sqrt{t} = t^{-1/2}$; $t = g(x) = 1 + x^2$ (с композицией $y = f(g(x))$). По правилу строки 3 таблицы производных (см. также примечание, помещенное после таблицы производных) и теореме о производной суммы функций, замечая, что производная от константы равна нулю, вычисляем производные:

$$f_t'(t) = (-1/2)t^{-3/2} \text{ и } g'(x) = 2x.$$

В результате имеем

$$f'(x) = (-1/2)(1 + x^2)^{-3/2}2x = -x(1 + x^2)^{-3/2}.$$

Математик вправе ожидать, что будет получен один и тот же результат при разных разложениях функции в композицию других функций. Разложим функцию из нашего примера в композицию трех функций: $f(t) = 1/t$, $t = h(s) = \sqrt{s}$, $s = g(x) = 1+x^2$ и итоговая композиция $y = f(h(g(x)))$. Естественно, что "трехступенчатая" производная считается по формуле

$$y' = f_t'(h(g(x)))\ h_s'(g(x))\ g'(x).$$

Вычислим сомножители: $f_t'(t) = (t^{-1})' = (-1)t^{-2}$, куда надо подставить $t = \sqrt{1 + x^2}$, в результате чего получается

$$(-1)\frac{1}{1 + x^2}.$$

Далее, $h_s' = (s^{-1/2})/2$ и после подстановки имеем второй сомножитель

$$\frac{1}{2\sqrt{1 + x^2}}.$$

И, наконец,

$$g'(x) = 2x.$$

Перемножая все три сомножителя, получаем

$$f'(x) = (-1)(1/2)(1/(1 + x^2)(1/\sqrt{1 + x^2})2x = -x(1 + x^2)^{-3/2}.$$

Результаты совпали.

Пример 2. Рассмотрим взаимно обратные функции $f(t)$ и $g(x)$, т.е. такие, что $f(g(x)) = x$ (например, $f(t) = \sin t$ и $g(x) = \arcsin x$).

По формуле производной сложной функции

$$(f(g(x)))' = f'(g(x))g'(x).$$

Поскольку $f(g(x)) = x$, эта производная равна единице. Это значит, что

$$f'(g(x)) = 1/g'(x).$$

Проверим это соотношение для функций $f(t) = \sin t$ и $g(x) = \arcsin x$. Мы должны сначала продифференцировать первую функцию: $f'(t) = \cos t$, после чего подставить $g(x)$ вместо t

$$f'(g(x)) = \cos(\arcsin x) = \sqrt{1 - \sin^2(\arcsin x)} = \sqrt{1 - x^2}.$$

В таблице находим

$$(\arcsin x)' = \frac{1}{\sqrt{1 - x^2}}$$

Соотношение $f'(g(x)) = 1/g'(x)$ выполнено[1].

Рис. 6.1 поясняет рассуждения. Для функций $f(x)$ и $g(x)$ выполнено $f(g(x)) = x$. Это значит, что если $g(a) = b$, то $f(b) = a$. Точки $(a;\, b)$ и $(b;\, a)$ симметричны относительно биссектрисы первого координатного угла (задаваемой уравнением $y = x$), поэтому симметричны будут и графики функций. Касательные в точке $(a;\, b)$ — к графику функции $f(x)$ и в точке $(b;\, a)$ — к графику $g(x)$ также симметричны. Это значит, что углы наклона касательных к оси абсцисс α и β в сумме дают 90 градусов. Для таких углов выполняется соотношение $\operatorname{tg}\alpha = 1/\operatorname{tg}\beta$.

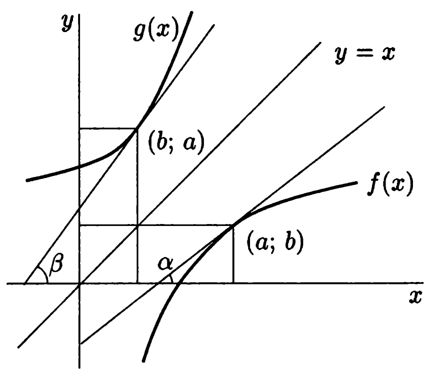


Рис. 6.1

Само собой разумеется, этому же соотношению подчиняются и производные функций $f(x)$ и $g(x)$ (поскольку производная — это тангенс

[1] На самом деле именно это соотношение служит для первоначального нахождения производных от функций $\arcsin x, \arccos x, \operatorname{arctg} x, \operatorname{arcctg} x$, а также e^x, после того как найдены производные $\sin x, \cos x, \operatorname{tg} x, \operatorname{ctg} x$ и $\ln x$.

угла наклона касательной к графику функции). Надо только быть внимательным: $f'(b) = 1/g'(a)$ — соотносятся производные в разных точках.

Но $b = g(a)$, поэтому $f'(g(a)) = 1/g'(a)$. Это и есть искомое соотношение.

6.2. Использование формулы производной сложной функции в неопределенном интеграле

Определение 6.1. *Неопределенный интеграл функции $f(x)$ обозначается*

$$\int f(x)\,dx$$

и представляет собой всю совокупность первообразных функции $f(x)$. Вычислить его — это значит найти одну из первообразных. После нахождения такой функции неопределенный интеграл записывается так:

$$\int f(x)\,dx = F(x) + C,$$

где C обозначает произвольную константу.

Пример 3. В таблицах интегралов элементарных функций мы не найдем интеграла

$$\int \cos x \sin x\,dx.$$

Формула производной сложной функции позволяет увидеть решение задачи: $\cos x$ есть производная от функции $\sin x$. Таким образом, если найти функцию $f(t)$, производная от которой будет равна t, то по правилу дифференцирования сложной функции $(f(\sin x))' = f'_t(\sin x)\cos x = \sin x \cos x$. Это и есть наша подынтегральная функция. Но такую функцию $f(t)$ найти очень легко — это функция $y = t^2/2$. Подставив вместо t в эту формулу $\sin x$, получаем $(\sin^2 x)/2$. Найдем для проверки производную от этой функции, используя наше новое правило:
$((\sin^2 x)/2)' = \sin x \cos x$, и мы получили подынтегральную функцию. Это значит, что мы действительно нашли первообразную в виде сложной функции.

Ответ задачи:

$$\int \cos x \sin x \, dx = (\sin^2 x)/2 + C.$$

Пример 4. Формула производной сложной функции может пребывать под интегралом в более скрытом виде. Пусть нам надо посчитать интеграл

$$\int \cos^3 x \, dx.$$

Чтобы сделать задачу похожей на предыдущую, преобразуем подынтегральную функцию: $\cos^3 x = (1 - \sin^2 x)\cos x = \cos x - \sin^2 x \cos x$. После этого разложим интеграл в разность двух интегралов.

$$\int \cos^3 x \, dx = \int \cos x \, dx - \int \sin^2 x \cos x \, dx.$$

Первый интеграл из суммы можно найти в таблицах, а второй похож на интеграл предыдущего примера, только вместо $t^2/2$ надо взять $t^3/3$, поскольку производная от $t^3/3$ равна t^2. Подставим $\sin x$ вместо t и получим

$$\int \cos^3 x \, dx = \sin x - \sin^3 x/3 + C.$$

Упражнение 6.1. Доказать, что если $F(x)$ первообразная для $f(x)$, $G(x)$ первообразная для $g(x)$, то $F(x) \pm G(x)$ первообразная для $f(x) \pm g(x)$, и обосновать этим законность разложения неопределенного интеграла в сумму (разность).

Для того чтобы легко справляться с задачами интегрирования, надо запомнить вид производных табличных функций и научиться распознавать их в сложных формулах.

Пример 5. Пусть нам надо посчитать интеграл

$$\int \frac{\sin x}{\cos^3 x} \, dx.$$

Здесь возможны даже два решения.

Во-первых, можно увидеть, что в числителе стоит производная от функции $y = -\cos x$, а в знаменателе степень того же $\cos x$. По таблице производных $(t^{-2}/(-2))' = t^{-3}$. Подставим $\cos x$ вместо t и получим

$$\left(\frac{1}{-2\cos^2 x}\right)' = (1/\cos^3 x)(-\sin x).$$

Это почти то, что надо, осталось только поменять знак:

$$\int \frac{\sin x}{\cos^3 x}\,dx = \frac{1}{2\cos^2 x} + C.$$

Второй способ.

Можно заметить, что

$$\frac{\sin x}{\cos^3 x} = \operatorname{tg} x \frac{1}{\cos^2 x},$$

т.е. подынтегральная функция раскладывается в произведение тангенса на его производную (см. таблицу). Взяв функции $f(t) = t^2/2$ и $g(x) = \operatorname{tg} x$, получим, что производная от $f(g(x))$ равна $t/\cos^2 x$, если вместо t подставить $\operatorname{tg} x$, т.е. равна как раз подынтегральной функции. Окончательный ответ:

$$\int \frac{\sin x}{\cos^3 x}\,dx = (\operatorname{tg}^2 x)/2 + C.$$

Упражнение 6.2. Доказать, что ответы, полученные двумя способами, совпадают (вспомнить для этого тригонометрическое тождество).

6.3. Замена переменной в неопределенном интеграле с использованием знака дифференциала

В выражении

$$\int \frac{\sin x}{\cos^3 x}\,dx$$

знак dx указывает на то, что x представляет собой независимую переменную и надо найти такую первообразную, производная которой именно по независимой переменной x равна подынтегральной функции.

Мы не будем здесь давать строгого определения дифференциала, принятого в современном анализе, а истолкуем его совершенно формально, как весьма удобный в задачах интегрирования технический символ.

Если $y = f(x)$ есть некоторая функция от x, то дифференциал y определяется формулой $dy = f'(x)\,dx$. Чтобы избежать путаницы, лучше написать формулу так:

$$dy = f'_x(x)\,dx$$

— производная берется по x.

Как видим, дифференциал функции определяется через дифференциал переменной dx. Если x, в свою очередь, является функцией переменной t и эта зависимость выражается формулой $x = g(t)$, то мы можем написать $dx = g'_t(t)\,dt$. Если, используя это равенство, подставить $g'_t(t)\,dt$ вместо dx в формулу $dy = f'_x(x)\,dx$, то получим $dy = f'_x(x)g'_t(t)dt$, и вместо x надо теперь подставить $x = g(t)$. Но то же самое выражение мы получим, если рассмотрим "сквозную" зависимость: y от t: $y = f(g(t))$.

В этом случае дифференциал y выразится формулой

$$dy = (f(g(x)))'\,dt = f'_x(g(t))\,g'_t(t)\,dt.$$

Формулы совпадают.

Положим $h(t) = f(g(t))$ и изобразим проведенные выше рассуждения в виде диаграммы:

$$\begin{array}{ccc} dy & = & f'(x)\,dx \\ \| & & \| \\ h'(t)\,dt & = & f'_x(g(t))\,g'_t(t)\,dt. \end{array}$$

Диаграмма интерпретируется следующим образом:

1) можно выразить дифференциал сложной функции dy (левый верхний угол диаграммы) сначала через дифференциал промежуточной переменной x (верхнее горизонтальное равенство), а затем подставить выражение dx через дифференциал независимой переменной dt (правое вертикальное равенство);

2) можно выразить дифференциал сложной функции dy сразу через производную композиции функций и дифференциал независимой переменной (левое вертикальное равенство), вычислив затем эту производную по правилу дифференцирования сложной функции (нижнее равенство);

— полученные разными путями результаты совпадут.

Припишем теперь к формулам диаграммы знак неопределенного интеграла слева. В левом верхнем углу окажется искомая функция $y = h(t)$.

$$\begin{array}{ccc} h(t) & = & \int f'(x)\,dx \\ \| & & \| \\ \int h'(t)\,dt & = & \int f'_x(g(t))\,g'_t(t)\,dt. \end{array}$$

Остальные интегралы поясним на примере.

Рассмотрим интеграл

$$\int \sin^3 t \cos t \, dt.$$

Ситуация задачи на диаграмме выглядит так:

$$\begin{array}{ccc} ? & = & ? \\ \| & & \| \\ ? & = & \int \sin^3 t \cos t \, dt. \end{array}$$

Чтобы решить задачу нахождения интеграла прямо, как мы делали в примерах предыдущего параграфа, надо в подынтегральной функции непосредственно распознать производную от неизвестной функции $h(t)$. Этому соответствует ход по диаграмме через нижнюю левую вершину.

Метод замены переменной позволяет решать задачу поэтапно, двигаясь через правую верхнюю вершину. Сначала мы видим, что $\cos t \, dt$ представляет собой дифференциал функции $x = \sin t$, поэтому, делая замену, получаем новый вариант диаграммы

$$\begin{array}{ccc} ? & = & \int x^3 \, dx \\ \| & & \| \\ * & = & \int \sin^3 t \cos t \, dt. \end{array}$$

Выражение $\int x^3 \, dx$ прямо указывает на другую функцию в искомой композиции: $y = x^4/4$. Таким образом, мы нашли сложную функцию $y = (\sin x)^4/4$, которая является одной из первообразных для подынтегральной функции $\sin^3 t \cos t$ и может быть помещена в левый верхний угол диаграммы:

$$\begin{array}{ccc} (\sin t)^4/4 & = & \int x^3 \, dx \\ \| & & \| \\ * & = & \int \sin^3 t \cos t \, dt. \end{array}$$

Пример 6. В третьей части книги, где разъясняются вопросы теории вероятностей, очень полезны оказываются интегралы, похожие на следующий:

$$\int x e^{-x^2} \, dx.$$

Вычислим его с помощью подстановки $t = -x^2$. В этом случае $x \, dx = -dt/2$, и интеграл приобретает вид

$$-1/2 \int e^t dt = -e^t/2 + C = -e^{-x^2}/2 + C.$$

Для проверки можно продифференцировать результат и убедиться, что

$$(-e^{-x^2}/2)' = xe^{-x^2}.$$

6.4. Интегрирование по частям

Мы уже знакомы с формулой производной от произведения двух функций:

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x).$$

Если равны функции, то и их интегралы будут равны:

$$\int (f(x)g(x))'dx = \int f'(x)g(x)dx + \int f(x)g'(x)dx.$$

Интеграл в левой части требует найти функцию, производная от которой равнялась бы $(f(x)g(x))'$. Эта задача тривиальна: само произведение $f(x)g(x)$, разумеется, является первообразной для своей производной. Мы можем теперь записать равенство в виде

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx.$$

Эту формулу называют формулой интегрирования по частям. Ее можно переписать в дифференциалах, заменив выражения $g'(x)dx$ на $d(g(x))$, а $f'(x)dx$ на $d(f(x))$:

$$\int f(x)d(g(x)) = f(x)g(x) - \int g(x)d(f(x)).$$

Когда эта формула дает эффект?

Пример 7. Вычислить интеграл

$$\int xe^x\,dx.$$

Положим $d(g(x)) = e^x dx$, а $f(x) = x$. Чтобы теперь воспользоваться формулой, нам надо вычислить первообразную для $y = e^x$ и продифференцировать $f(x) = x$. Обе задачи достаточно просты: поскольку $(e^x)' = e^x$, то $g(x) = e^x$. Действительно, $d(e^x) = e^x dx$. Еще проще вторая задача: $d(f(x)) = d(x) = dx$.

При решении задач удобно выписать последние выкладки в виде таблицы, в которой первая строка заполняется сразу, как только выбраны "части", а вторая — после необходимых вычислений.

$$\begin{array}{rr} d(g(x)) = e^x dx & f(x) = x \\ g(x) = e^x & d(f(x)) = d(x) = dx \end{array}.$$

Теперь вместо

$$\int xe^x dx = \int f(x)d(g(x))$$

мы будем считать

$$f(x)g(x) - \int g(x)d(f(x)) = xe^x - \int e^x dx.$$

Последний интеграл равен $e^x + C$. Таким образом, окончательный ответ:

$$\int xe^x dx = xe^x - e^x + C.$$

В этом примере "части" подбираются так, что дифференцирование одной из них ее упрощает, а нахождение первообразной для другой "части" ее не усложняет.

Пример 8. Вычислить интеграл

$$\int 2x \ln x \, dx.$$

Положим $d(g(x)) = 2x\,dx$, а $f(x) = \ln x$. Чтобы воспользоваться формулой, надо вычислить первообразную для $y = 2x$ и продифференцировать $f(x) = \ln x$. Поскольку $(x^2)' = 2x$, то $g(x) = x^2$. Далее, $d(f(x)) = d(\ln x) = dx/x$. Таблица приобретает следующий вид:

$$\begin{array}{rr} d(g(x)) = 2x\,dx & f(x) = \ln x \\ g(x) = x^2 & d(f(x)) = d(\ln x) = dx/x. \end{array}$$

Вместо

$$\int 2x \ln x \, dx = \int f(x)d(g(x))$$

считаем

$$f(x)g(x) - \int g(x)d(f(x)) = x^2 \ln x - \int \frac{x^2 dx}{x}.$$

Сокращая x в числителе и знаменателе, находим, что последний интеграл равен $x^2/2 + C$. Таким образом, окончательный ответ:

$$\int 2x \ln x \, dx = x^2 \ln x - x^2/2 + C.$$

В этом примере эффект достигается благодаря тому, что усложнение первой "части" не так велико, а упрощение второй "части" кардинально.

Глава 7

Функции и интегралы в бесконечных пределах

7.1. Поведение функций на бесконечности

В конце третьей главы мы рассматривали один частный случай интегралов с бесконечными пределами интегрирования. Прежде чем продолжить разговор об этих, как их называют, несобственных интегралах, мы обобщим наше понимание предела функции на те случаи, когда вместо конечных чисел приходится иметь дело с бесконечностью.

В первой главе было дано "литературное определение предела": число A является пределом функции $f(x)$ при x стремящемся к x_0, если можно обеспечить как угодно малое отличие $f(x)$ от A, если только выбрать достаточно близкое к x_0 значение аргумента x.

Теперь нам понадобятся пределы функций при x стремящемся к ∞. Определение предела для этого случая мало чем отличается от данного выше — надо заменить x_0 на ∞ и понять, что значит "выбрать достаточно близкое к ∞ значение аргумента x". Приближаться к ∞ означает становиться все больше и больше, поэтому определение приобретает следующий вид:

Число A является пределом функции $f(x)$ при x стремящемся к ∞, если можно обеспечить как угодно малое отличие $f(x)$ от A, если только выбрать достаточно большое значение аргумента x.

Упражнение 7.1. Дать определение предела функции при x стремящемся к $-\infty$.

Пример 1. Если $n \geq 1$, то

$$\lim_{x\to\infty} \frac{1}{x^n} = 0.$$

Для того чтобы обеспечить неравенство

$$\frac{1}{x^n} < 0{,}001$$

достаточно взять x больше 1000.

Упражнение 7.2. Убедиться, что

$$\lim_{x\to\infty} \frac{1}{\sqrt[n]{x}} = 0.$$

Насколько большим надо взять x, чтобы эта функция стала меньше, чем 0,001?

Теоремы о пределах, упомянутые и доказанные в предыдущих главах, остаются верными и для случая стремления x к $\pm\infty$: предел суммы и произведения функций равен соответственно сумме и произведению пределов; предел частного двух функций равен частному пределов этих функций, если только предел знаменателя отличен от нуля.

Еще одна модификация определения также будет полезна в дальнейшем: функция $f(x)$ стремится к ∞ (к $-\infty$) при x стремящемся к ∞, если можно обеспечить как угодно большое (соответственно, как угодно большое по модулю, отрицательное) значение $f(x)$, если только выбрать достаточно большое значение аргумента x.

Предел суммы и предел произведения двух функций, имеющих предел ∞, также равен бесконечности, но предел частного может принимать самые разные значения.

Пример 2. Найти

$$\lim_{x\to\infty} \frac{x^4 + x^2}{x^4 + 1}.$$

Числитель и знаменатель дроби стремятся к бесконечности, но предел их отношения конечное число.

Чтобы доказать это, поделим числитель и знаменатель на x^4, после чего применим теоремы о пределах частного и суммы, поскольку все рассматриваемые пределы конечны.

$$\lim_{x\to\infty} \frac{x^4+x^2}{x^4+1} = \lim_{x\to\infty} \frac{1+\frac{1}{x^2}}{1+\frac{1}{x^4}} = \frac{\lim_{x\to\infty}(1+\frac{1}{x^2})}{\lim_{x\to\infty}(1+\frac{1}{x^4})} = \frac{1+\lim_{x\to\infty}\frac{1}{x^2}}{1+\lim_{x\to\infty}\frac{1}{x^4}} = \frac{1+0}{1+0} = 1.$$

Пример 3. Найти

$$\lim_{x\to\infty} \frac{x^3 + x^2}{x^4 + 1}.$$

Снова числитель и знаменатель дроби стремятся к бесконечности, но в данном случае предел их отношения равен нулю. Опять поделим числитель и знаменатель на x^4

$$\lim_{x\to\infty} \frac{x^3 + x^2}{x^4 + 1} = \lim_{x\to\infty} \frac{\frac{1}{x} + \frac{1}{x^2}}{1 + \frac{1}{x^4}} = \frac{\lim_{x\to\infty}(\frac{1}{x} + \frac{1}{x^2})}{\lim_{x\to\infty}(1 + \frac{1}{x^4})} = \frac{0+0}{1+0} = 0.$$

В приведенных примерах был дан метод вычисления предела отношения многочленов при x стремящемся к ∞. Для вычисления такого предела надо числитель и знаменатель поделить на старшую степень x. Если степени числителя и знаменателя равны, то предел будет равен отношению старших коэффициентов, если не равны, то предел равен нулю, когда степень знаменателя больше степени числителя, и бесконечности (со знаком "плюс" или "минус") в противоположном случае.

Упражнение 7.3. Объяснить, почему

1) если $f(x) > 0$ и $\lim\limits_{x\to\infty} f(x) = 0$, то $\lim\limits_{x\to\infty} \dfrac{1}{f(x)} = \infty$;

2) если $\lim\limits_{x\to\infty} f(x) = \infty$, то $\lim\limits_{x\to\infty} \dfrac{1}{f(x)} = 0$.

Рассмотрим функцию $f(x) = e^x/x$ при $x > 1$. Эта функция возрастает, поскольку

$$f'(x) = \frac{e^x x - e^x}{x^2} = \frac{e^x}{x} - \frac{e^x}{x^2},$$

причем ее производная всегда больше единицы[1], т.е. график функции в каждой точке поднимается к бесконечности круче, чем проходящая через эту точку прямая, образующая с осью абсцисс угол 45^o. Это позволяет нам заключить, что

$$\lim_{x\to\infty} \frac{e^x}{x} = \infty.$$

Если считать упражнение 7.3 выполненным, то из предыдущего следует, что

$$\lim_{x\to\infty} \frac{x}{e^x} = \lim_{x\to\infty} xe^{-x} = 0.$$

[1] Для того чтобы в этом убедиться, можно посчитать несколько первых значений функции $2^x/x - 2^x/x^2$ при целых значениях аргумента $x = 4, 5, \ldots$ и заметить, что $e^x > 2^x$, а значит, и $e^x/x - e^x/x^2 > 2^x/x - 2^x/x^2$.

На математическом жаргоне можно сказать так: e^x стремится к бесконечности быстрее, чем x.

Верно даже более сильное утверждение: e^{x^α} стремится к бесконечности быстрее, чем x^n, как бы ни был мал показатель степени $\alpha > 0$ и как бы ни был велик показатель n (имеется в виду, что оба они постоянные числа):

$$\lim_{x\to\infty} \frac{e^{x^\alpha}}{x^n} = \infty.$$

Например

$$\lim_{x\to\infty} \frac{e^{\sqrt[100]{x}}}{x^{1000}} = \infty,$$

т.е. $e^{\sqrt[100]{x}}$ стремится к бесконечности быстрее, чем x^{1000}.

Похожим образом,

$$\lim_{x\to\infty} \frac{x}{\ln x} = \infty,$$

и даже

$$\lim_{x\to\infty} \frac{\sqrt[100]{x}}{\ln x^{1000000}} = \infty.$$

Рассматривая пределы функций при стремлении переменной к $-\infty$, надо иметь в виду, что

$$\lim_{x\to-\infty} e^x = 0.$$

Вследствие этого и приведенных выше формул

$$\lim_{x\to-\infty} xe^x = 0.$$

7.2. Правило Лопиталя

Для того чтобы найти предел отношения $f(x)/g(x)$, если обе функции стремятся при этом к нулю или обе функции стремятся к бесконечности, часто оказывается полезным применить так называемое правило Лопиталя, которое мы приведем в двух из четырех возможных вариантов и притом вовсе без доказательства.

1) Если

$$\lim_{x\to\infty} f(x) = 0 \text{ и } \lim_{x\to\infty} g(x) = 0,$$

то

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}.$$

Разумеется, применять это правило можно лишь в случае, если производные и предел в правой части существуют.

2) Если

$$\lim_{x\to\infty} f(x) = \infty \text{ и } \lim_{x\to\infty} g(x) = \infty,$$

то

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)},$$

если существуют производные и предел в правой части

Упражнение 7.4. Используя правило Лопиталя, показать, что

$$(1)\ \lim_{x\to\infty} \frac{e^x}{x} = \infty, \quad (2)\ \lim_{x\to\infty} \frac{\ln(x^{20})}{x} = 0, \quad (3)\ \lim_{x\to-\infty} xe^x = 0.$$

7.3. Интегралы с бесконечными пределами интегрирования

Введем новое обозначение. Если $F(x)$ первообразная для $f(x)$, то иногда удобно записывать в формуле Ньютона—Лейбница промежуточный результат:

$$\int_a^b f(x)\,dx = F(x)\ \Big|_a^b = F(b) - F(a).$$

Выражение $F(x)\ \Big|_a^b$ указывает на то, что в первообразную $F(x)$ надо подставить пределы интегрирования и взять соответствующую разность.

В третьей главе мы упоминали о том, что в теории вероятностей важную роль играют интегралы по бесконечным промежуткам, так называемые несобственные интегралы.

Пример 4.

$$\int_1^\infty \frac{1}{x^2}\,dx$$

Вычислим сначала интеграл в конечных пределах от единицы до некоторого переменного t. Первообразная для функции $y = 1/x^2$ это функ-

ция $y = -1/x$, и наш интеграл вычисляется подстановкой в эту первообразную верхнего предела интегрирования t и нижнего 1:

$$\int_1^t \frac{1}{x^2}\,dx = -1/x\ \Big|_1^t = \left(-\frac{1}{t}\right) - \left(-\frac{1}{1}\right) = 1 - \frac{1}{t}.$$

Какое бы большое значение t мы ни взяли, т.е. как бы ни был велик промежуток интегрирования, значение интеграла всегда остается меньшим единицы, приближаясь к единице при неограниченном возрастании t. Мы можем в таком случае считать, что

$$\int_1^\infty \frac{1}{x^2}\,dx = 1.$$

Для того чтобы площадь бесконечной криволинейной трапеции была конечной, как в предыдущем примере, необходимо, чтобы подынтегральная функция стремилась к нулю. Но этого недостаточно.

Пример 5. Рассмотрим площадь под стандартной гиперболой:

$$\int_1^\infty \frac{1}{x}\,dx$$

Как и в предыдущем случае, вычислим интеграл в конечных пределах от единицы до некоторого переменного t. Первообразная для функции $y = 1/x$ это функция $y = \ln x$, подставляя в эту первообразную верхний t и нижний 1 пределы интегрирования, получаем:

$$\int_1^t \frac{1}{x}\,dx = \ln x\ \Big|_1^t = \ln t.$$

Поскольку

$$\lim_{t\to\infty} \ln t = \infty,$$

то значение интеграла бесконечно растет по мере возрастания верхнего предела интегрирования t. Мы можем сказать, что площадь под стандартной гиперболой бесконечна.

Пример 6.

$$\int_0^\infty xe^{-x}\,dx.$$

Непосредственно вычислив производную, можно убедиться, что функция $y = -xe^{-x} - e^{-x}$ является первообразной для подынтегральной функции $y = xe^{-x}$ (найти ее можно интегрированием по частям — см. главу 6), поэтому

$$\int\limits_0^t xe^{-x}\,d = (-xe^{-x} - e^{-x})\Big|_0^t = (-te^{-t} - e^{-t}) + 0e^{-0} + e^{-0} = 1 - te^{-t} - e^{-t}.$$

При неограниченном увеличении t

$$\lim_{t\to\infty}(1 - te^{-t} - e^{-t}) = 1 - \lim_{t\to\infty} te^{-t} - \lim_{t\to\infty} e^{-t}.$$

Последний предел очевидно равен нулю, а предпоследний — предел функции $y = te^{-t}$ — был разобран в предыдущем параграфе и также равен нулю. Таким образом,

$$\int\limits_0^\infty xe^{-x}\,dx = 1.$$

Глава 8

Одно приложение идеи дифференциала: закон Вебера—Фехнера

8.1. Дифференциал как приращение

В предыдущих главах мы уже использовали равенство

$$dy = f'(x)dx,$$

понимая его чисто формально. Теперь мы рассмотрим понятие дифференциала более пристально. На рис. 8.1 изображена функция $f(x)$, ее касательная в точке x_0 и некоторая точка x, лежащая вблизи x_0.

Разность значений $f(x)-f(x_0)$, или, как говорят, приращение функции, измеряется отрезком BD, который состоит из двух частей — BC и CD. BC представляет собой линейную часть приращения, которая определяется тангенсом угла наклона касательной, т.е. производной $f'(x_0)$, умноженной на приращение аргумента $x-x_0$. Из чертежа понятно, что чем меньше x отличается от x_0, тем меньшую долю приращения составляет нелинейная добавка CD в общем приращении BD. Поэтому выражение

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

можно заменить на эквивалентное по смыслу выражение

$$f(x) - f(x_0) = f'(x_0)(x - x_0) + \alpha(x)(x - x_0),$$

где

$$\lim_{x \to x_0} \alpha(x) = 0.$$

Если обозначить $f(x) - f(x_0) = \Delta y$ и $x - x_0 = \Delta x$, то последнее равенство можно переписать в таком виде:

$$\Delta y = f'(x)\Delta x + \alpha(x)\Delta x.$$

На рис. 8.1 $\alpha(x)$ это отношение CD к AB, которое стремится к нулю при x стремящемся к x_0. Вместо

$$\Delta y = f'(x)\Delta x + \alpha(x)\Delta x, \text{ при } \lim_{x \to x_0} \alpha(x) = 0.$$

пишут

$$dy = f'(x)dx.$$

Во времена Лейбница, который ввел понятие дифференциала и используемую до сих пор символику, говорили, что бесконечно малое приращение функции равно (в точности равно) бесконечно малому приращению аргумента[1], умноженному на производную $f'(x)$. Последнее равенство можно переписать в виде

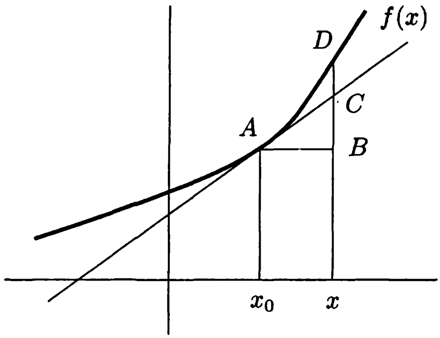


Рис. 8.1

$$\frac{dy}{dx} = f'(x),$$

что означает в терминах Лейбница, что производная есть отношение бесконечно малого приращения функции к бесконечно малому приращению аргумента.

Мы будем понимать равенства $dy = f'(x)dx$ и $\frac{dy}{dx} = f'(x)$ как пропорции если не бесконечно малых, то очень маленьких приращений — настолько маленьких, что равенство становится почти точным.

[1] В современной математике умеют обходиться без этих не вполне ясных бесконечно малых и определяют дифференциал как главную линейную часть приращения функции.

8.2. Закон Вебера—Фехнера

В 1834 году Э. Вебер опубликовал результаты своих опытов, которые легли в основание целой отрасли психологии — психофизики. Ее разработка, как и название, принадлежат Г. Фехнеру, который через четверть века продолжил опыты Вебера и дал им математическую интерпретацию.

Э. Вебер обнаружил, что минимально воспринимаемая разница в весе между грузами пропорциональна самим сравниваемым весам. Человек в состоянии отличить груз 62 г от груза 60 г, но не может отличить 62 г от 61 г и 61 от 60. Если предложить сравнить грузы, весящие около 120 г, то картина будет такова: 120 уверенно отличается от 124, но не от 123 г.

В дальнейшем процедуры, служившие Э. Веберу и Г. Фехнеру для расчетов этих так называемых дифференциальных порогов чувствительности, подверглись существенному уточнению. В самом деле, понятно, что различение грузов или каких-то иных стимулов не является строго детерминированным процессом и 120 будет иногда справедливо охарактеризовано как меньший вес по отношению к 123 г, но в каких-то случаях испытуемый будет настолько неточен, что сочтет груз 120 г более тяжелым, чем 121 г и т.д.

Эти очень интересные вопросы о способах корректного измерения дифференциальных порогов мы здесь не будем рассматривать, как не будем обсуждать некоторые важные отклонения от закона Вебера—Фехнера вблизи абсолютных порогов ощущений, т.е. вблизи границ, вне которых раздражитель вообще не воспринимается. Мы примем на веру следующее утверждение: при измерении стимуляции в некоторых естественных физических единицах (для грузов это вес, для звуковых и световых раздражителей это интенсивность звука и света) дифференциальные пороги чувствительности, которые характеризуются равной частотой обнаружения различия стимулов, пропорциональны величине стимула.

В случае веса дифференциальный порог, или минимально воспринимаемый прирост веса Δr, будет составлять примерно 1/30 от веса стимула r, т.е. при любой величине r

$$\frac{\Delta r}{r} = \frac{1}{30}.$$

Фехнер предположил далее, что одинаковые частоты обнаружения различий стимулов при разлных величинах их физической меры объ-

ясняются тем, что им соответствуют равные приросты ощущения. Т.е. прибавление к тестовому весу 1/30 его величины увеличивает субъективное ощущение на некоторую фиксированную величину:

$$\Delta S = k\frac{\Delta r}{r}.$$

Константа k говорит о том, что мы пока свободны выбирать единицы измерения ощущений.

Не останавливаясь на этом, Фехнер предположил, что опыты с дифференциальными порогами указывают на более общую закономерность, чем связь между дифференциальным порогом Δr и соответствующим именно ему приростом ощущения ΔS. Последняя формула выражает связь между всяким приростом величины стимула и соответствующим приростом ощущения.

Для математика вполне понятно, что это отношение тем точнее, чем меньше прирост величины стимула, поэтому закономерность можно выразить формулой

$$dS = k\frac{dr}{r}$$

или

$$\frac{dS}{dr} = k\frac{1}{r}.$$

Если $S(r)$ функция, выражающая ощущение S через величину стимула r, то последние формулы означают, что $S'(r) = k\frac{1}{r}$, откуда

$$S(r) = k\ln r + C.$$

Теперь мы вольны выбрать единицы измерения ощущений и тем определить константы k и C. Фехнер ограничился следующим рассуждением: если считать, что нулевое ощущение соответствует тому минимальному стимулу r_0, который вообще воспринимается (нижний абсолютный порог ощущения), то

$$0 = k\ln r_0 + C \quad \text{и} \quad C = -k\ln r_0.$$

Отсюда $S = k(\ln r - \ln r_0) = k\ln(r/r_0)$. Окончательно

$$S = k\ln\frac{r}{r_0},$$

или в словесной форме: величина ощущения пропорциональна логарифму величины стимула.

Последняя формулировка и называется законом Вебера—Фехнера.

Часть III

Теория вероятностей

Глава 1

Случайные события и вероятности

1.1. Различные подходы к понятию вероятности

Понятие случайного события является основополагающим в изучении вероятностных методов и моделей. Под *случайным событием* будем понимать событие, которое может произойти или не произойти в результате некоторого *испытания*. При этом испытанием может быть как целенаправленное действие, так и явление, происходящее независимо от наблюдателя. В дальнейшем случайные события будем называть просто *событиями*.

Приведем несколько примеров.

Пример 1. Испытание — бросается монета. Возможные события — выпадение "герба" или "цифры".

Пример 2. Наступает день 12 января — испытание. "В течение дня наблюдается ясная погода" — событие.

Пример 3. Студент сдает экзамен — испытание. "Он получил оценку 5" — событие.

Каждому событию может быть поставлено в соответствие число, принадлежащее отрезку $[0, 1]$ и называемое *вероятностью* данного события. Вероятность можно понимать как меру достоверности (в том числе и субъективной) данного события. В таком смысле слово "вероятность" употребляется и в бытовой речи, где, однако, ее обычно "измеряют" в процентах — от 0 до 100%. Вероятность обычно обозначают

буквой p (от англ. probability — вероятность). Чем более достоверным представляется наступление события, тем больше его вероятность. Вероятность невозможного события считается равной нулю, вероятность абсолютно достоверного события считается равной единице.

Для определения вероятностей событий возможны различные подходы.

Начнем с рассмотрения ситуации, когда в результате испытания может произойти один из некоторого конечного множества *равновозможных исходов* (*пространства исходов*). Если общее число исходов (или, иначе говоря, элементарных событий) равно n, то каждому из них приписывается вероятность $1/n$.

Пример 4. Бросается игральный кубик, на гранях которого нанесено разное число точек — от 1 до 6 включительно (рис. 1.1). Тогда исходов будет шесть: "выпало число 1", "выпало число 2", ..., "выпало число 6". Коротко пространство исходов можно записать следующим образом:

$$\{1,2,3,4,5,6\}.$$

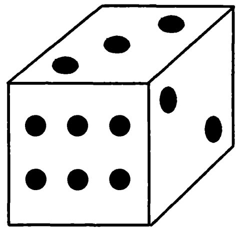


Рис. 1.1. Кубик с пронумерованными гранями

Вероятность выпадения каждого из этих чисел равна 1/6 (как говорят, "один шанс из шести").

Событием можно считать любое подмножество пространства исходов. И обратно, любое событие является подмножеством пространства исходов. Будем говорить, что событие A произошло, если результат (исход) испытания принадлежит множеству A. (Здесь и далее события будем обозначать, как правило, прописными латинскими буквами.) Продолжая пример 4, можно заметить, что событию A_1 = "выпало четное число очков" соответствует подмножество {2,4,6} пространства исходов, а событию A_2 = "выпало число очков, большее двух" соответствует подмножество {3,4,5,6}.

Посмотрим теперь на ситуацию с более общей точки зрения.

Классическая вероятность

Пусть n — число всех равновозможных исходов, а m — число исходов, составляющих событие A. Вероятность события A (обозначение $p(A)$) определяется следующим образом

$$p(A) = \frac{m}{n}.$$

Это так называемая *классическая вероятность*. В частности, для упомянутых выше событий A_1 и A_2 имеем

$$p(A_1) = \frac{1}{2}, \quad p(A_2) = \frac{2}{3}.$$

Подчеркнем, что формула классической вероятности предполагает конечность числа исходов n. Обратимся теперь к случаю, когда число исходов бесконечно.

Геометрическая вероятность

Пусть на плоскости имеется фигура F, содержащая фигуру f (рис. 1.2). Испытание заключается в том, что в фигуру F наугад бросается точка. Тем самым пространство исходов можно отождествить с этой фигурой. Здесь число исходов бесконечно (у фигуры F бесконечно много точек), притом все исходы имеют одинаковые шансы осуществиться. Определим A как событие, заключающееся в том, что брошенная точка попала в фигуру f. Тогда вероятность события A (*геометрическая вероятность*) определяется следующим образом

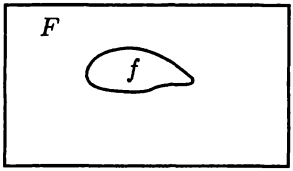


Рис. 1.2

$$p(A) = \frac{S_f}{S_F},$$

где S_F и S_f — площади фигур F и f соответственно.

Аналогично определяется геометрическая вероятность на прямой и в пространстве, только вместо площадей фигур в формуле для вероятности надо поставить соответственно длины и объемы.

Пример 5. В результате урагана был оборван телефонный кабель между 20-м и 60-м километрами линии. Какова вероятность того, что обрыв произошел между 30-м и 35-м километрами?

Здесь $l_F = 60 - 20 = 40$, а $l_f = 35 - 30 = 5$. Значит, $p = 5/40 = 1/8$.

Статистическая вероятность

Предположим, что событие A может произойти либо не произойти в результате некоторого эксперимента. Повторим эксперимент n раз и подсчитаем, сколько раз произошло событие A. Пусть это число равно m. Отношение m/n назовем *относительной частотой появления*

события A в n испытаниях. Если при достаточно больших значениях n относительные частоты группируются около некоторой постоянной, то эту постоянную будем считать *статистической вероятностью* события A

$$p(A) \approx \frac{m}{n} \qquad \text{при больших } n.$$

Пример 6. Если подбросить монету n раз и подсчитать число m выпадений герба, то при достаточно большом n отношение m/n будет близко к 0,5 (если монета симметричная — не гнутая, не смещен центр тяжести и пр.)

Субъективная вероятность

Во многих реальных ситуациях определение вероятности событий одним из приведенных выше способов невозможно. Тогда на первый план выступает отмеченное выше понимание вероятности как меры достоверности того или иного события. В этом случае следует провести экспертный опрос и на основе его результатов получить *субъективную вероятность* события.

Пример 7. Какова вероятность того, что некто станет президентом на ближайших выборах? Ясно, что здесь может идти речь о вероятности только в субъективном смысле.

Замечание. С принятием некоторого числа в качестве субъективной вероятности связаны два достаточно независимых действия. Во-первых, требуется правильно провести опрос и, во-вторых, надо правильно учесть уже высказанное мнение экспертов. При этом возникает ряд психологических и математических проблем. Их обсуждение, однако, выходит за рамки этой книги.

1.2. Формулы алгебры событий. Несовместимые и независимые события

Если *определены* вероятности элементарных событий, можно переходить к *вычислению* вероятностей более сложных событий, являющихся комбинацией определенных ранее элементарных.

Предположим, что с некоторым испытанием связаны события A и B. Их *суммой* назовем событие, заключающееся в том, что произошло хотя бы одно из событий — A или B (обозначение: $A + B$).

Пример 8. Пусть A = "это случилось в сентябре...", B = "это случилось в октябре...", C = "это случилось в ноябре...". Тогда $(A + B + C)$ = "это случилось осенью...".

Произведением событий A и B назовем событие, состоящее в совместном наступлении этих событий (обозначение AB).

Пример 9. Пусть A = "в аудиторию вошел студент", B = "в аудиторию вошел человек в темных очках". Тогда AB = "в аудиторию вошел студент в темных очках".

Событием *противоположным* A назовем событие, состоящее в том, что A не произошло (обозначение: $\bar{A}$, "не A").

Пример 10. Пусть испытанием является бросок баскетболиста по кольцу, A = "баскетболист попал". Тогда $\bar{A}$ = "баскетболист не попал".

Введенные понятия допускают простую геометрическую интерпретацию. Рассуждения в рамках этой интерпретации хотя и не являются доказательствами в строгом смысле, но вполне достаточны для понимания предмета. Пусть испытанием является бросание точки в прямоугольную область на плоскости, обозначенную на рисунках буквой Ω, а событиями A, B, C и D — попадание точки в области, которые мы обозначим теми же буквами — A, B, C и D соответственно. Тогда сумма событий A и B заштрихована на рис. 1.3, сумма $C + D$ — на рис. 1.4. Произведение AB показано на рис. 1.5, произведение CD —

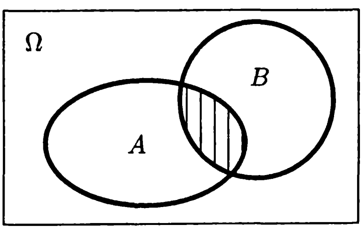


Рис. 1.3

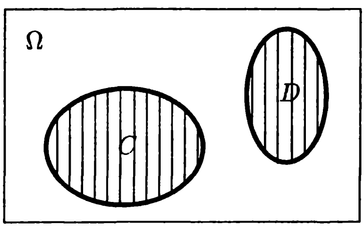


Рис. 1.4

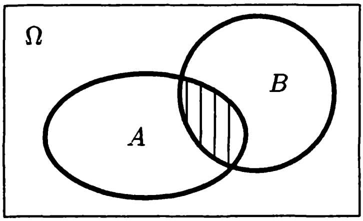


Рис. 1.5

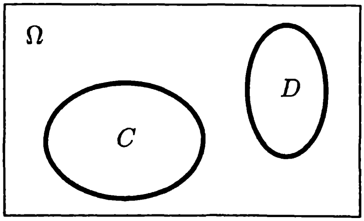


Рис. 1.6

на рис. 1.6, событие противоположное A — на рис. 1.7. Заметим, что произведение CD является невозможным событием, $CD = \emptyset$.

Перейдем теперь к вычислению вероятностей событий $A+B$, AB и $\bar{A}$, считая известными вероятности событий A и B.

Вероятность события $\bar{A}$ вычисляется легко:

$$p(\bar{A}) = 1 - p(A).$$

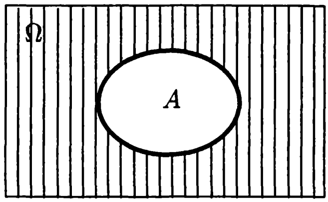


Рис. 1.7

Для вероятности события $A + B$ справедлива следующая формула

$$p(A + B) = p(A) + p(B) - p(AB). \tag{1}$$

В это соотношение входит пока неизвестная нам вероятность произведения AB. Впрочем, часто слагаемое $p(AB)$ оказывается равным 0. Рассмотрим эту ситуацию подробнее.

Если события A и B не могут произойти одновременно в результате одного испытания (иными словами, если AB — невозможное событие), то их называют *несовместимыми*, и тогда $p(AB) = 0$. Если же события могут произойти в результате одного испытания, то их называют *совместимыми*.

Пример 11. События A и $\bar{A}$ несовместимы.

Пример 12. События A и B на рис. 1.3, 1.5 совместимы.

Пример 13. События C и D на рис. 1.4, 1.6 несовместимы.

Для случая несовместимых событий формула (1) приобретает особенно простой вид

$$p(A + B) = p(A) + p(B). \tag{2}$$

В психологии вопрос о зависимости различных характеристик исследуемого процесса возникает очень часто. Например, зависит ли оценка студента по математике от его пола? Зависит ли результат теста интеллекта подростка от его показателей по тесту исследовательской активности в детском возрасте? Разберем вопрос о независимости событий на простых моделях.

В случаях физической независмости событий их вероятности перемножаются. Два последовательных подбрасывания монеты явно дают независимые результаты. Это значит, что возможные результаты этого двойного испытания можно записать в виде таблицы

$$\begin{matrix} \text{ГГ} & \text{ГЦ} \\ \text{ЦГ} & \text{ЦЦ}, \end{matrix}$$

где ГГ обозначает последовательное выпадение двух "гербов", ГЦ выпадение сначала "герба", потом "цифры", ЦГ — "цифры" и "герба", ЦЦ — двух "цифр".

Все четыре исхода в силу симметрии равновероятны, поэтому $p(\text{ГГ}) = 1/4 = p(\text{Г})p(\text{Г})$. Вообще, для физически независимых событий A и B

$$p(AB) = p(A)\,p(B).$$

Верно также и

$$p(\bar{A}B) = p(\bar{A})\,p(B);\; p(A\bar{B}) = p(A)\,p(\bar{B});\; p(\bar{A}\bar{B}) = p(\bar{A})\,p(\bar{B}).$$

Аналогичные формулы верны и для большего числа независимых событий. Например

$$p(ABC) = p(A)\,p(B)\,p(C)$$

для независимых A, B и C.

Пример 14. Четыре стрелка одновременно стреляют по цели. Вероятности попадания в цель для каждого стрелка известны: 0,7; 0,75; 0,7 и 0,65 соответственно. Чему равна вероятность того, что цель будет поражена (хотя бы одним стрелком)?

Решение. Обозначим за A_i $(i = 1, 2, 3, 4)$ событие, состоящее в том, что i-й стрелок попал в цель. Эти события независимы, их вероятности по условию таковы

$$p(A_1) = 0{,}7; \qquad p(A_3) = 0{,}7;$$
$$p(A_2) = 0{,}75; \qquad p(A_3) = 0{,}65.$$

Цель не будет поражена (событие $\bar{A}$), если все стрелки промахнутся

$$\bar{A} = \bar{A}_1 \bar{A}_2 \bar{A}_3 \bar{A}_4.$$

Вычисляя вероятность, получаем

$$p(A) = 1 - p(\bar{A}_1)\, p(\bar{A}_2)\, p(\bar{A}_3)\, p(\bar{A}_4) = \\ = 1 - 0{,}3 \cdot 0{,}25 \cdot 0{,}3 \cdot 0{,}35 = 0{,}992125.$$

1.3. Вычисление вероятностей

Перейдем к рассмотрению важного вопроса: как вычислять вероятности сложных событий, если известны вероятности простых. Подчеркнем еще раз, что вероятности простых событий определяются предварительно в классическом, геометрическом либо субъективном понимании.

Единого алгоритма решения произвольной вероятностной задачи не существует. Рассмотрим два взаимодополняющих метода — применение формул (*"аналитический"* метод) и применение дерева вероятностей (*"графический"* метод). Рассмотрение будем вести на примерах.

Пример 15. Известно, что в среднем 5% изделий некоторой фирмы бракованные. Взяли наугад на проверку два изделия. Какова вероятность того, что ровно одно из этих двух изделий будет забраковано?

Решение 1. Обозначим за $Б_1$ ($Б_2$) событие, состоящее в том, что первое (второе) изделие оказалось бракованным. Тогда $\bar{Б}_1$ ($\bar{Б}_2$) — противоположное событие, состоящее в том, что первое (второе) изделие удовлетворяет стандартным требованиям качества. Интересующее нас событие A = {"ровно одна деталь бракована"} можно представить следующим образом: A = {"первая деталь бракована" *и* "вторая деталь не бракована" *или* "первая деталь не бракована" *и* "вторая деталь бракована"}. Вспомнив, что логическим *и, или, не* соответствуют в формулах алгебры событий умножение, сложение, противоположное событие, запишем

$$A = Б_1 \bar{Б}_2 + \bar{Б}_1 Б_2.$$

Теперь перейдем к вычислению вероятности события A. Заметим, что:

1) события $Б_1 \bar{Б}_2$ и $\bar{Б}_1 Б_2$ несовместимы (они не могут наступить одновременно);

2) события $Б_1$ и $\bar{Б}_2$, а также $\bar{Б}_1$ и $Б_2$ независимы.

Поэтому

$$p(A) = p(\text{Б}_1\bar{\text{Б}}_2) + p(\bar{\text{Б}}_1\text{Б}_2) = p(\text{Б}_1)p(\bar{\text{Б}}_2) + p(\bar{\text{Б}}_1)p(\text{Б}_2). \qquad (8)$$

Теперь осталось подставить вероятности "простых" событий Б_1 и Б_2. По условию 5% изделий бракованы. Поэтому

$$p(\text{Б}_1) = p(\text{Б}_2) = 0{,}05;$$
$$p(\bar{\text{Б}}_1) = p(\bar{\text{Б}}_2) = 1 - 0{,}05 = 0{,}95.$$

Подставляя в формулу (8), получаем

$$p(A) = 0{,}05 \cdot 0{,}95 + 0{,}95 \cdot 0{,}05 = 0{,}095.$$

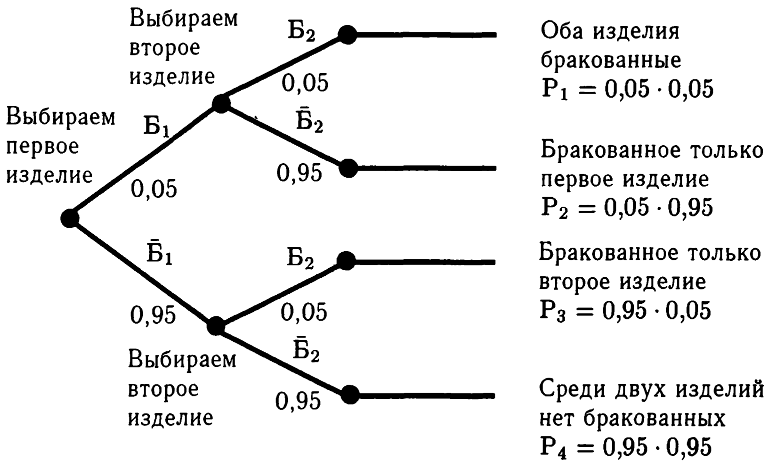


Рис. 1.8

Решение 2. Построим так называемое *дерево вероятностей*, учитывающее все возможные исходы (рис. 1.8). Здесь вершинам дерева (кроме концевых) соответствуют испытания, а ребрам — события. Сначала рассмотрим первое изделие. При этом возможны два исхода — изделие может оказаться бракованным (с вероятностью 0,05) либо качественным (с вероятностью 0,95). В каждом из этих случаев рассмотрим второе изделие, которое тоже может быть либо бракованным, либо качественным (с теми же вероятностями).

В результате получаем четыре возможности, обозначаемые концевыми вершинами дерева. К каждой из этих возможностей ведет путь

из начальной точки, состоящий из двух ребер дерева. Для нахождения вероятностей p_1, p_2, p_3, p_4 перемножаются вероятности ребер соответствующего пути.

Искомая вероятность вычисляется как сумма p_2 и p_3:

$$p_2 + p_3 = 0{,}095.$$

Замечание 1. Поскольку все возможные исходы в сумме составляют достоверное событие, то суммарная вероятность всегда равна единице. В данном случае

$$p_1 + p_2 + p_3 + p_4 = 1.$$

Пример 16. Через остановку пролегают троллейбусный и автобусный маршруты. Троллейбус подъезжает через каждые 15 минут, автобус — через каждые 25 минут. К остановке подходит пассажир. Какова вероятность того, что в ближайшие 10 минут на остановке появится троллейбус либо автобус?

Решение 1. Пассажир подошел к остановке в некоторый случайный момент между двумя последовательными приездами троллейбуса. По условию троллейбус подъезжает через каждые 15 минут. По формуле геометрической вероятности найдем вероятность $p(T)$ того, что троллейбус появится на остановке в ближайшие 10 минут:

$$p(T) = \frac{10}{15} = \frac{2}{3}.$$

Вероятность $p(A)$ того, что в ближайшие 10 минут на остановку подъедет автобус, такова

$$p(A) = \frac{10}{25} = \frac{2}{5}.$$

Пассажир не уедет с остановки в ближайшие 10 минут, если не приедут ни троллейбус, ни автобус, то есть если произойдет событие $\bar{T}\bar{A}$. События $\bar{T}$ и $\bar{A}$ независимы, поэтому

$$p(\bar{T}\bar{A}) = p(\bar{T})p(\bar{A}) = \frac{1}{3} \cdot \frac{3}{5} = \frac{1}{5} = 0{,}2.$$

Вероятность же того, что пассажир уедет, составляет

$$1 - 0{,}2 = 0{,}8.$$

Замечание 2. Еще один способ рассуждений состоит в применении формулы (1) на стр. 209

$$p(S) = p(T + A) = p(T) + P(A) - p(TA) = \frac{2}{3} + \frac{2}{5} - \frac{2}{3} \cdot \frac{2}{5} =$$

$$= \frac{16}{15} - \frac{4}{15} = \frac{12}{15} = 0{,}8.$$

(Разумеется, события T и A совместимы — могут подъехать и троллейбус, и автобус.)

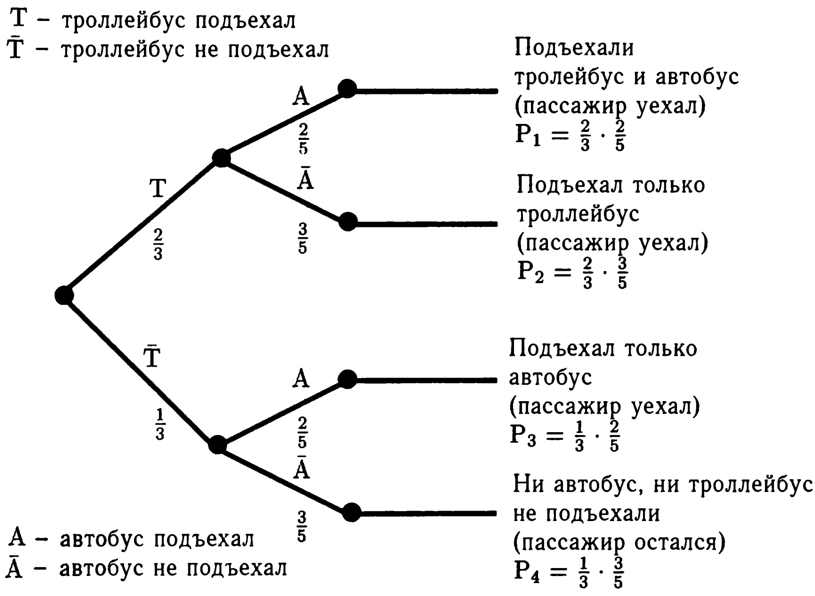


Рис. 1.9

Решение 2. Построим дерево вероятностей (рис. 1.9). Интересующая нас вероятность вычисляется как сумма вероятностей попарно несовместимых событий

$$p_1 + p_2 + p_3 = \frac{2}{3} \cdot \frac{2}{5} + \frac{2}{3} \cdot \frac{3}{5} + \frac{1}{3} \cdot \frac{2}{5} = \frac{12}{15} = 0{,}8.$$

Для расчетов вероятностей в случаях, когда события не являются независимыми, вводится понятие *условной вероятности*. Услов-

ная вероятность события A, если произошло событие B, обозначается $p(A|B)$, (читается "вероятность A при условии B").

Для любых событий A и B (как независимых, так и зависимых) справедлива следующая формула

$$p(AB) = p(A|B)p(B). \tag{3}$$

Если события независимы, то $p(A|B) = p(A)$, поэтому

$$p(AB) = p(A)p(B). \tag{4}$$

Пример 17. Студент пришел на зачет, зная 15 вопросов из 20. Если студент не может ответить, ему предоставляется еще одна (но не более!) попытка. Какова вероятность сдать зачет?

Решение 1. Введем следующие обозначения:

A_1 — студент сразу вытянул знакомый билет (и сдал зачет);

$\bar{A}_1$ — студент вытянул незнакомый билет (еще одна попытка);

A_2 - студент со второго раза наконец-то вытянул знакомый билет (и сдал зачет);

$\bar{A}_2$ — студент и во второй раз вытянул незнакомый билет (и ему предстоит пересдача);

A — студент сдал зачет.

Студент сдает зачет, если он либо сразу вытянул знакомый билет, либо вытянул сначала незнакомый, а во второй раз — знакомый билет. Формально это можно записать следующим образом

$$A = A_1 + \bar{A}_1 A_2.$$

Переходя к вероятностям, получаем

$$p(A) = p(A_1) + p(\bar{A}_1)p(A_2|\bar{A}_1) = \frac{15}{20} + \frac{5}{20} \cdot \frac{15}{19} = \frac{18}{19}.$$

Здесь условную вероятность вычисляем прямо: поскольку во второй попытке "участвует" уже 19 билетов и из них по-прежнему 15 знакомых, то $p(A_2|\bar{A}_1)$ равна 15/19.

Замечание. Можно было рассуждать иначе. Студент не сдает зачет, если и в первый, и во второй раз вытянет незнакомый билет:

$$\bar{A} = \bar{A}_1 \bar{A}_2.$$

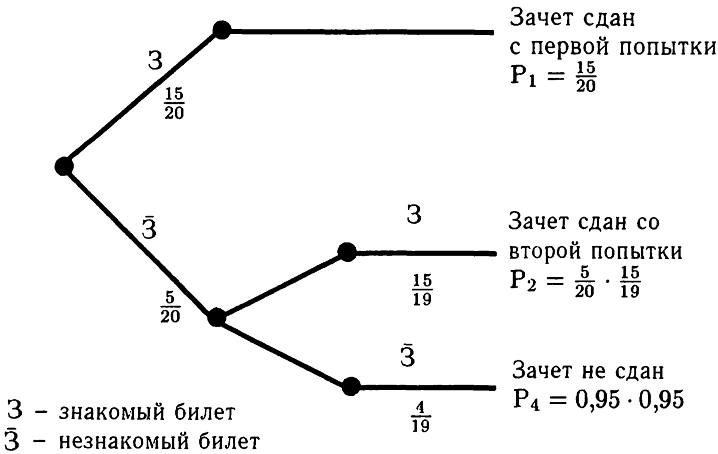


Рис. 1.10

Поэтому, поскольку после первой неудачной попытки остается только 4 незнакомых билета из 19, то

$$p(\bar{A}) = p(\bar{A}_1)p(\bar{A}_2\,|\,\bar{A}_1) = \frac{5}{20}\cdot\frac{4}{19} = \frac{1}{19}.$$

Отсюда

$$p(A) = 1 - \frac{1}{19} = \frac{18}{19}.$$

Решение 2. Построим дерево вероятностей (рис. 1.10). Из него легко получить ответ:

$$p_1 + p_2 = \frac{15}{20} + \frac{5}{20}\cdot\frac{15}{19} = \frac{18}{19}.$$

Пример 18. Игроки A и B разыгрывают денежный приз в следующей игре. Подбрасывается монета до тех пор, пока не выпадет шесть "гербов" либо шесть "цифр". Если выпало шесть "гербов", то выигрывает игрок A, если шесть "цифр" — игрок B. Монету подбросили 8 раз. При счете 5:3 в пользу игрока A (то есть выпало пять "гербов" и три "цифры") игра прервалась по независящим от игроков причинам. В каком отношении надо поделить денежный приз?

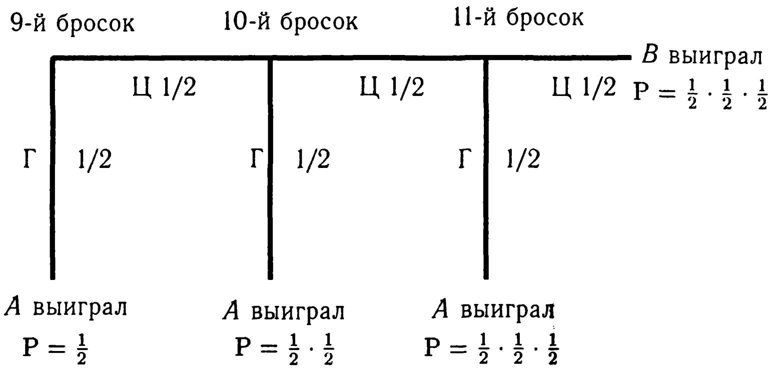


Рис. 1.11

Решение 1. Если бы игра продолжалась, ситуация могла бы развиваться (начиная с девятого подбрасывания монеты) следующим образом (приведем все возможные варианты и их вероятности):

Г — игрок *А* выиграл, вероятность 1/2;

Ц Г — игрок *А* выиграл, вероятность 1/4;

Ц Ц Г — игрок *А* выиграл, вероятность 1/8;

Ц Ц Ц — игрок *В* выиграл, вероятность 1/8;

Таким образом, при счете 5:3 вероятность выигрыша игрока А составляет

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}=\frac{7}{8},$$

а игрока *В* — всего 1/8. По-видимому, приз следует разделить в отношении 7:1 в пользу игрока *А*.

Замечание. Можно было и не перебирать все возможные варианты, а просто заметить, что игрок *В* выигрывает лишь в случае выпадения трех цифр подряд. Вероятность этого составляет 1/8, а вероятность выигрыша игрока *А* (что является противоположным событием, ведь ничья правилами игры не предусмотрена) составляет соответственно

$$1-\frac{1}{8}=\frac{7}{8}.$$

Решение 2. Дерево вероятностей см. на рис. 1.11. Вывод тот же, что и в *решении 1*.

Глава 2

Формула полной вероятности и формула Байеса

2.1. Формула полной вероятности

Одним из эффективных методов подсчета вероятностей является формула полной вероятности, являющаяся следствием формул для вероятностей суммы и произведения событий.

Пример 1. Предположим, что 5% всех мужчин и 0,25% всех женщин страдают дальтонизмом. Для простоты будем считать, что мужчин и женщин одинаковое число. Какова вероятность того, что наугад выбранное лицо страдает дальтонизмом?

Решение. Построим дерево вероятностей (рис. 2.1). Справа обозначены лишь два исхода из четырех возможных, поскольку оставшиеся два нас в данном случае не интересуют. Из рисунка видно, что вероятность того, что наугад выбранное лицо дальтоник, составляет

$$p_1 + p_2 = 0{,}5 \cdot 0{,}05 + 0{,}5 \cdot 0{,}0025 = 0{,}02625.$$

Последнюю формулу можно обосновать без использования дерева вероятностей, и в некоторых случаях это создает определенные удобства.

Предположим, что событие A может наступить только вместе с одним из попарно несовместимых событий $H_1, \ldots, H_n$ (по отношению к

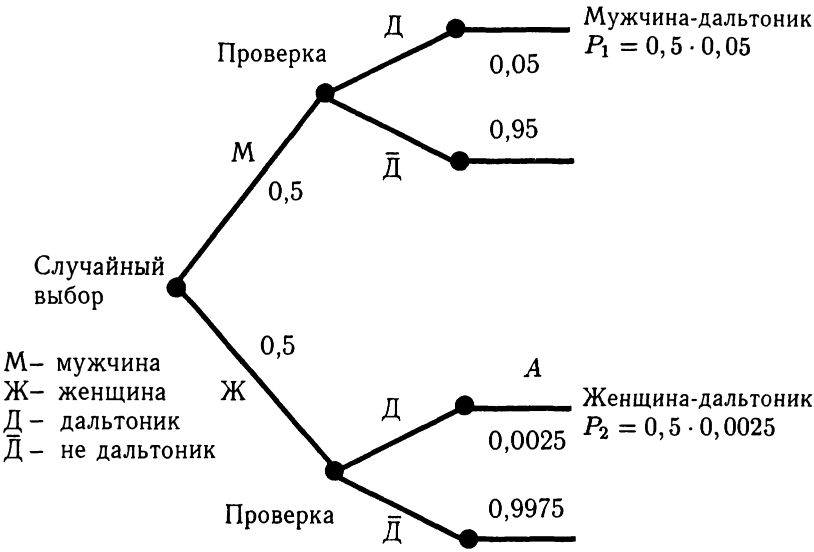


Рис. 2.1

событию A будем называть их *гипотезами* — в предыдущем примере гипотезы "выбранное лицо мужчина" и "выбранное лицо женщина"). Тогда появление события A связано с обязательным появлением ровно одного из событий $AH_1, \ldots, AH_n$ и A можно представить в виде

$$A = AH_1 + \ldots + AH_n$$

(см. рис. 2.2, где $n = 3$).

Пример 2. Пусть в доме пять дверей. Событие A = "человек вошел в дом", гипотеза H_i = "человек прошел через i-ю дверь", где $i = 1, 2, 3, 4, 5$.

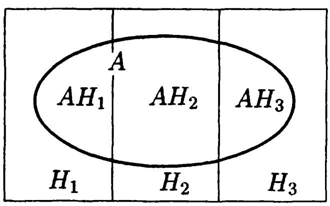


Рис. 2.2

Поскольку события $H_1, \ldots, H_n$ попарно несовместимы, то таковыми же будут и события $AH_1, \ldots, AH_n$ (это легко понять из рис. 2.2 — поскольку несовместимость H_1 и, например, H_2 означает отсутствие области, общей H_1 и H_2, то и AH_1 и AH_2 не могут иметь общих точек).

Поскольку, как мы помним из предыдущей главы, вероятность суммы несовместимых событий равна сумме вероятностей событий-слагаемых, то

$$p(A) = p(AH_1 + \ldots + AH_n) = p(AH_1) + \ldots + p(AH_n).$$

Наконец, вспоминая еще одну формулу предыдущей главы: $p(AB) = p(A \mid B)p(B)$, выражающую вероятность произведения событий через условную вероятность, переписываем последнее равенство в виде

$$p(A) = p(A \mid H_1)p(H_1) + \ldots + p(A \mid H_n)p(H_n).$$

Это и есть *формула полной вероятности*, которую можно использовать вместо дерева вероятностей при подсчете вероятностей конкретных событий.

Вернемся к примеру 1 и используем теперь новую формулу.

Решение 2. Пусть гипотеза H_1 = "выбранное лицо — мужчина", H_2 = "выбранное лицо — женщина", A = "выбранное лицо страдает дальтонизмом". Требуется вычислить вероятности $p(A)$. Имеем

$$p(H_1) = p(H_2) = 0,5, \; p(A \mid H_1) = 0,05, \; p(A \mid H_2) = 0,0025.$$

Тогда по формуле полной вероятности

$$p(A) = 0,05 \cdot 0,5 + 0,0025 \cdot 0,5 = 0,02625.$$

2.2. Формула Байеса

С формулой полной вероятности тесно связана формула Байеса. Еще раз вернемся к примеру 1. Пусть наугад выбранное лицо страдает дальтонизмом. Какова в таком случае вероятность того, что это мужчина?

Прежде чем считать вероятность, разберемся, какой смысл имеет эта, так называемая апостериорная вероятность.

Если сотни и тысячи раз повторять опыт с проверкой на дальтонизм случайно выбранных испытуемых и отмечать, с какой частотой среди выявленных дальтоников встречаются мужчины и женщины, то эта частота будет статистическим приближением к искомой вероятности. Эту вероятность можно вычислить, не прибегая к дорогостоящим экспериментам.

Пусть опыт произведен и наступило событие A. Напомним, что как и в предыдущем параграфе, событие A могло произойти только вместе с одной из гипотез $H_1, \ldots, H_n$. Поэтому можно вычислить вероятность

того, что имело место именно событие H_i. Эта *апостериорная* вероятность $p(H_i \mid A)$ отличается, вообще говоря, от априорной вероятности $p(H_i)$ (в которой не учтен тот факт, что событие A произошло).

Записывая формулы для вероятности произведения событий, имеем

$$p(AH_i) = p(A \mid H_i)p(H_i),$$

$$p(AH_i) = p(H_i \mid A)p(A).$$

Приравняв правые части последних формул, получаем равенство

$$p(A \mid H_i)p(H_i) = p(H_i \mid A)p(A).$$

И далее

$$p(H_i \mid A) = \frac{p(A \mid H_i)p(H_i)}{p(A)}.$$

Привлекая формулу полной вероятности, получаем в итоге *формулу Байеса*:

$$p(H_i \mid A) = \frac{p(A \mid H_i)p(H_i)}{p(A \mid H_1)p(H_1) + \ldots + p(A \mid H_n)p(H_n)}.$$

Пример 3. Предположим, что в двух корзинах содержится соответственно 3 белых и 7 черных шаров и 7 белых и 3 черных шара. Наугад выбирают корзину и из нее наугад вынимают шар. Этот шар оказывается белым. Какова вероятность того, что была выбрана корзина с большим числом белых шаров?

Решение. Здесь H_1 = “выбрана первая корзина”, H_2 = “выбрана вторая корзина”, A = “вынутый шар оказался белым”. Требуется вычислить вероятность $p(H_2 \mid A)$.

Имеем

$$p(H_1) = p(H_2) = \frac{1}{2}, \; p(A \mid H_1) = \frac{3}{10}, \; p(A \mid H_2) = \frac{7}{10},$$

по формуле Байеса

$$p(H_2 \mid A) = \frac{p(A \mid H_2)p(H_2)}{p(A \mid H_1)p(H_1) + p(A \mid H_2)p(H_2)} = \frac{\frac{7}{10} \cdot \frac{1}{2}}{\frac{3}{10} \cdot \frac{1}{2} + \frac{7}{10} \cdot \frac{1}{2}} = 0{,}7\,.$$

Пример 4. На экзамене студентам предлагается 20 билетов, 5 из которых легкие, а 15 — трудные. Два студента по очереди тянут билеты — сначала первый студент, затем второй.

а) Чему равна вероятность вытянуть легкий билет для первого студента?

б) Чему равна вероятность вытянуть легкий билет для второго студента?

в) Известно, что второй студент вытянул легкий билет. Чему равна вероятность того, что и первый вытянул легкий?

Решение. Введем обозначения:

H_1 — первый студент вытянул легкий билет;

H_2 — первый студент вытянул трудный билет;

A — второй студент вытянул легкий билет.

Тогда ответ на вопрос пункта а) дает формула классической вероятности

$$p(H_1) = \frac{5}{20} = \frac{1}{4},$$

ответ на вопрос пункта б) формула полной вероятности

$$p(A) = p(A|H_1)\,p(H_1) + p(A|H_2)\,p(H_2) = \frac{4}{19}\cdot\frac{5}{20} + \frac{5}{19}\cdot\frac{15}{20} = \frac{1}{4},$$

а ответ на вопрос пункта в) формула Байеса

$$p(H_1|A) = \frac{p(A|H_1)\,p(H_1)}{p(A)} = \frac{\frac{4}{19}\cdot\frac{5}{20}}{\frac{4}{19}\cdot\frac{5}{20} + \frac{5}{19}\cdot\frac{15}{20}} = \frac{4}{19}.$$

Пример 5. Фирма планирует выпуск на рынок нового вида товара. Субъективные представления руководства фирмы таковы: вероятность хорошего спроса на этот товар составляет 0,7, вероятность плохого спроса — 0,3. Было проведено специальное исследование товарного рынка, которое предсказало плохой сбыт. Однако известно, что исследования такого рода дают правильный прогноз не всегда, а лишь с вероятностью 0,8. Каким образом маркетинговое исследование повлияло на вероятности хорошего и плохого сбыта?

Решение. Введем следующие обозначения:

H_1 — сбыт будет хорошим;

H_2 — сбыт будет плохим;

A — исследование рынка предсказало плохой сбыт.

Опыт и интуиция руководства фирмы дают, в соответствии с условием, следующие вероятности

$$p(H_1) = 0{,}7,$$
$$p(H_2) = 0{,}3.$$

Маркетинговое исследование дает верный результат с вероятностью 0,8, поэтому

$$p(A|H_1) = 0{,}2,$$
$$p(A|H_2) = 0{,}8.$$

Подставляя все эти вероятности в формулу Байеса, получаем

$$p(H_1|A) = \frac{p(A|H_1)\,p(H_1)}{p(A|H_1)\,p(H_1)) + p(A|H_2)\,p(H_2)} = \frac{0{,}2 \cdot 0{,}7}{0{,}2 \cdot 0{,}7 + 0{,}3 \cdot 0{,}8} = \\ = \frac{7}{19} \approx 0{,}37,$$

$$p(H_2|A) = \frac{p(A|H_2)\,p(H_2)}{p(A|H_1)\,p(H_1)) + p(A|H_2)\,p(H_2)} = \frac{0{,}3 \cdot 0{,}8}{0{,}2 \cdot 0{,}7 + 0{,}3 \cdot 0{,}8} = \\ = \frac{12}{19} \approx 0{,}63.$$

Ответ: в результате исследования вероятность хорошего сбыта уменьшилась до 0,37, а вероятность плохого увеличилась до 0,63.

Глава 3

Схема испытаний Бернулли

Пусть A — случайное событие, которое может произойти в результате некоторого испытания. Допустим далее, что нас интересует лишь то, наступило ли событие A: будем считать возможными лишь два события — A и $\bar{A}$. Обозначим их вероятности через p и q соответственно, $p+q=1$.

Предположим, что испытание повторяется при одних и тех же условиях некоторое фиксированное количество раз, скажем, три раза. Построим дерево вероятностей (рис. 3.1) и вычислим вероятности каждого из восьми возможных событий:

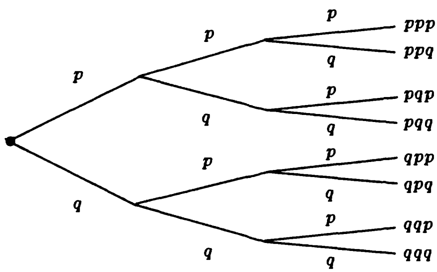


Рис. 3.1

$$
\begin{aligned}
p(AAA) &= ppp = p^3, & p(A\bar{A}\bar{A}) &= pqq = pq^2,\\
p(AA\bar{A}) &= ppq = p^2q, & p(\bar{A}A\bar{A}) &= qpq = pq^2,\\
p(A\bar{A}A) &= pqp = p^2q, & p(\bar{A}\bar{A}A) &= qqp = pq^2,\\
p(\bar{A}AA) &= qpp = p^2q, & p(\bar{A}\bar{A}\bar{A}) &= qqq = q^3.
\end{aligned}
$$

(запись $\bar{A}\bar{A}A$ обозначает событие "в первых двух испытаниях событие А не произошло, в третьем — произошло", аналогично записаны остальные 7 событий).

Как мы видим, вероятность каждого из исходов представима в виде $p^k q^{3-k}$, где k показывает число наступлений события A, а $(3 - k)$ соответственно число его ненаступлений.

Вероятностные схемы такого рода называются *схемами Бернулли*, или *схемами биномиальных экспериментов*. Эти схемы широко применяются при анализе реальных ситуаций в тех случаях, когда эксперимент можно считать *биномиальным*, т.е. когда

- он состоит из фиксированного числа n испытаний,
- в каждом из этих испытаний происходит либо не происходит некоторое событие,
- вероятность этого события одинакова в каждом испытании,
- испытания независимы одно от другого.

Пример 1. Тренированный стрелок совершает пять выстрелов по мишени, причем все выстрелы производятся практически в одних и тех же условиях. При этом число попаданий в "десятку" может меняться от 0 до 5.

Пример 2. В помёте, состоящем из 8 мышей, происходящих от одних родителей, число мышей, имеющих прямую, а не волнистую шерстку может равняться произвольному целому числу от 0 до 8.

Пример 3. Один за другим бросают три игральных кубика. Число выпадений "шестерки" может принимать одно из четырех значений от 0 до 3 включительно.

Вероятность того, что событие A, которое наступает при одном испытании с вероятностью p, произойдет ровно k раз после n испытаний, обозначим через $P(p, n, k)$ (ясно, что $0 \leqslant k \leqslant n$).

Справедлива следующая формула

$$P(p, n, k) = C_n^k p^k q^{n-k}.$$

Здесь p — это вероятность появления события A в одном испытании, $q = 1 - p$, а C_n^k (читается "цэ из эн по ка") называется *биномиальным коэффициентом* и вычисляется по любой из формул

$$C_n^k = \frac{n!}{k!(n-k)!},$$

$$C_n^k = \frac{n(n-1)\dots(n-k+1)}{k!},$$

$$C_n^k = \frac{n(n-1)\dots(k+1)}{(n-k)!},$$

где $n!$ (читается "эн факториал") — произведение натуральных чисел от 1 до n включительно.

$$n! = 1 \cdot 2 \cdot 3 \ldots (n-2)(n-1)n.$$

Заметим также, что по определению принимается

$$0! = 1.$$

Пример 4. Монету бросают 10 раз. Какова вероятность того, что при этом "герб" выпадет ровно 3 раза? Выпадет меньше двух раз?

Решение. Здесь $n = 10$, $p = q = 1/2$.

$$P\{\text{"герб" выпал ровно три раза}\} =$$
$$= P\left(\frac{1}{2}, 10, 3\right) = C_{10}^{3}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{7} = \frac{10\cdot 9\cdot 8}{1\cdot 2\cdot 3}\,\frac{1}{2^{10}} = \frac{15}{128}.$$

$$P\{\text{"герб" выпал меньше двух раз}\} =$$
$$P\{\text{"герб" не выпал ни разу}\} + P\{\text{"герб" выпал ровно один раз}\} =$$
$$= P\left(\frac{1}{2}, 10, 0\right) + P\left(\frac{1}{2}, 10, 1\right) = C_{10}^{0}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{10} + C_{10}^{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{9} =$$
$$= \frac{1}{2^{10}} + \frac{10}{2^{10}} = \frac{11}{1024}.$$

Пример 5. Студент пишет контрольную работу по теории вероятностей. У него есть предположение о том, как решить задачу, однако свою способность найти правильное решение студент оценивает невысоко — примерно 0,4.

Вокруг студента в аудитории сидят пять однокурсников. Можно рискнуть опросить их и принять либо отвергнуть решение на основании большинства голосов. Подготовку этих однокурсников студент оценивает так же, как и свою.

Как лучше поступить студенту — положиться на свои соображения или на большинство голосов однокурсников?

Решение. Для выбора между двумя альтернативами следует сначала выбрать какой-либо критерий. По-видимому, в данной ситуации таким критерием является вероятность правильно решить задачу. Опираясь на свои соображения студент получает вероятность 0,4.

Вычислим теперь вероятность того, что большинство из 5 опрошенных однокурсников даст правильный ответ. Большинство — это либо 3, либо 4, либо 5. Поэтому искомая вероятность вычисляется следующим образом:

$$P(0{,}4;\ 5;\ 3) + P(0{,}4;\ 5;\ 4) + P(0{,}4;\ 5;\ 5) =$$

$$= C_5^3 \cdot (0{,}4)^3 \cdot (0{,}6)^2 + C_5^4 \cdot (0{,}4)^4 \cdot 0{,}6 + C_5^5 \cdot (0{,}4)^5 =$$

$$= 10 \cdot 0{,}064 \cdot 0{,}36 + 5 \cdot 0{,}0256 \cdot 0{,}6 + 0{,}01024 = 0{,}31744$$

Вероятность снизилось с 0,4 до 0,31744 — более чем на 20%. Вывод: опрос однокурсников в данной ситуации лучше не проводить.

Упражнения к главам 1–3

Упражнение 1. На плоскости нанесена сетка квадратов со стороной 10 см. Найдите вероятность того, что брошенный на плоскость круг радиуса 1 см не пересечет стороны ни одного из квадратов.

Ответ: 0,64.

Упражнение 2. Имеются две сумки с мячами, в каждой по 5 мячей, пронумерованных от 1 до 5. Наугад вынимается по одному мячу из каждой сумки. Какова вероятность того, что это будут мячи с номерами 2 и 5 (безразлично, какой из них из какой сумки вынут)?

Ответ: 0,08.

Упражнение 3. Пусть испытанием является бросание точки в единичный квадрат, а событием A, B, C, D — попадание точки в соответствующую прямоугольную область (см. рис. 3.2, 3.3). Проверить совместимость и зависимость а) событий A и B (рис. 3.2) б)событий C и D (рис. 3.3).

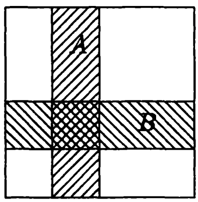


Рис. 3.2

Ответ: а) события A и B совместимы и независимы; б) события C и D несовместимы и зависимы.

Упражнение 4. Вероятность того, что в течение одной смены возникнет неполадка, равна 0,05. Какова вероятность того, что не произойдет ни одной неполадки за три смены?

Ответ: 0,857375.

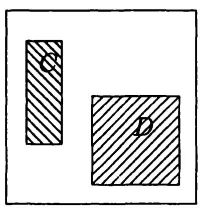


Рис. 3.3

Упражнение 5. Студент пришел на зачет, зная из 30 вопросов только 24. Какова вероятность сдать зачет, если после отказа отвечать на вопрос преподаватель задает еще один вопрос?

Ответ: $\frac{28}{29} \approx 0,9655$.

Упражнение 6. Два охотника стреляют в волка, причем каждый делает по одному выстрелу. Для первого охотника вероятность попадания в цель 0,7, для второго — 0,8. Какова вероятность попадания в волка (хотя бы при одном выстреле)? Как изменится результат, если охотники сделают по два выстрела?

Ответ: 0,94; 0,9964.

Упражнение 7. Из большой связки галстуков, в которой галстуки зеленого, красного и желтого цветов находятся в пропорции 5:3:2, двое мужчин случайным образом выбирают по одному галстуку. Какова вероятность того, что они выберут галстуки одинакового цвета?

Ответ: 0,38.

Упражнение 8. Имеются две урны. В первой находится 1 белый шар, 3 черных и 4 красных, во второй — 3 белых, 2 черных и 3 красных. Из каждой урны извлекают по шару. Найти вероятность того, что цвета вытащенных шаров совпадут.

Ответ: $\frac{21}{64} \approx 0,328$.

Упражнение 9. На трех станках различных марок изготовляется определенная деталь. Производительность 1-го станка за смену составляет 40 деталей, 2-го — 35 деталей, 3-го — 25 деталей. Установлено, что 2%, 3% и 5% продукции этих станков соответственно имеют скрытые дефекты. В конце смены на контроль взята одна деталь.

а) Какова вероятность того, что эта деталь нестандартная?

б) Если деталь оказалась нестандартной, какова вероятность того, что она изготовлена на первом, втором, третьем станке?

Ответ: а) 0,031; б) $\frac{8}{31} \approx 0,258$; $\frac{21}{62} \approx 0,339$; $\frac{25}{62} \approx 0,403$.

Упражнение 10. Два стрелка независимо один от другого делают по одному выстрелу по мишени. Вероятность попадания в мишень для первого стрелка 0,8, для второго — 0,4. После стрельбы в мишени обнаружена одна пробоина. Найти вероятность того, что в мишень попал первый стрелок.

Ответ: 6/7.

Упражнение 11. Как изменились бы вероятности хорошего и плохого сбыта в примере 5 на стр. 22, если бы исследование рынка предсказало хороший сбыт?

Ответ: вероятность хорошего сбыта — 0,9; вероятность плохого сбыта — 0,1.

Упражнение 12. Игральный кубик бросают 5 раз. Найдите вероятность того, что число очков, кратное трем, появится ровно два раза.

Ответ: 80/243.

Упражнение 13. Всхожесть семян растений данного сорта оценивается вероятностью, равной 0,8. Какова вероятность того, что из пяти посеянных семян взойдут не менее четырех?

Ответ: 0,73728.

Глава 4

Комбинаторика. Бином Ньютона

В предыдущей главе были введены биномиальные коэффициенты C_n^k, используемые для вычисления вероятностей событий в схеме испытаний Бернулли. Мы также использовали эти коэффициенты в первой главе второй части для вычисления производной от степенной функции. Теперь мы восполним пробел и выведем использовавшиеся формулы.

4.1. Размещения

При расчете вероятностей по "классическому" определению мы используем отношение количества благоприятных исходов к общему количеству возможных исходов. Здесь мы научимся считать эти количества исходов, как бы велики они ни были.

Наш первый модельный пример, с которым будут соотноситься другие примеры, таков:

Пример 1. Сколькими способами можно разложить k пронумерованных шаров в n пронумерованных корзин ($n \geq k$), так, чтобы в каждой корзине оказалось не больше одного шара.

Решим задачу сначала для $n = 4$, $k = 2$. Всего возможно 12 различных размещений (рис. 4.1, справа). Следующее рассуждение показывает, почему вариантов именно 12.

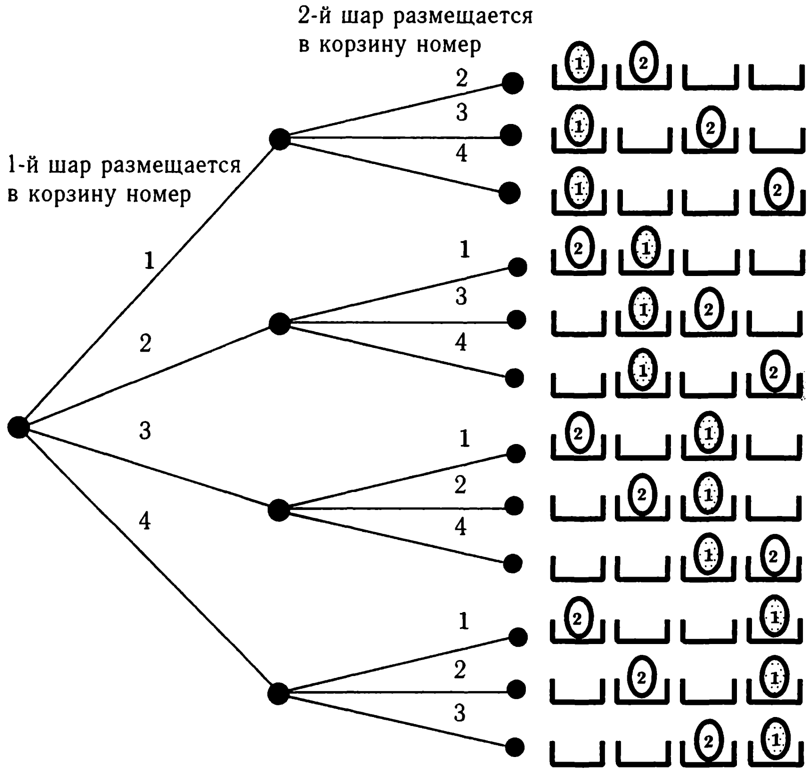


Рис. 4.1

Первый шар мы можем положить в любую из четырех имеющихся корзин, после чего второй шар может быть размещен в любой из оставшихся трех корзин. Можно представить выбор в виде дерева, каждая ветка которого оканчивается одним из вариантов размещения (рис. 4.1, слева).

Это рассуждение легко распространяется на случай произвольных n и k:
первый шар может быть положен в любую из n корзин,
второй шар — в любую из оставшихся $n-1$ корзин,
третий шар — в любую из оставшихся $n-2$ корзин,
. . .
k-й шар в любую из оставшихся $n-(k-1)$ корзин.

Всего получается $n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-(k-1))$ размещений.

Количество размещений k элементов в n ячеек обозначается A_n^k. Мы вывели формулу числа размещений:

$$A_n^k = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-(k-1)).$$

Если $n = k$, то последняя строка будет оканчиваться словами "в любую из одной корзины". Этот последний частный случай размещений называется перестановками совокупности n элементов. Количество перестановок n элементов равно в точности $n!$.

4.2. Сочетания

Следующий вопрос, который вплотную подводит нас к формуле бинома Ньютона, звучит так: сколькими способами можно выбрать из n различимых предметов k штук?

Пример 2. Сколькими способами можно выбрать из четырех пронумерованных корзин две?

Мы свяжем этот пример с предыдущим следующим образом: выбранные корзины будем отмечать тем, что положим в них шары. Тот же рис. 4.1 дает первый шаг решения — в первой строке обозначен выбор первой и второй корзины, во второй строке — первой и третьей и т.д.

Однако можно заметить, что каждый выбор пары корзин встречается в списке из двенадцати размещений дважды. Первый выбор можно найти также в четвертой строке, второй выбор — в седьмой и т.д. В случае размещений для нас существенно, какой шар оказался в данной корзине, в случае сочетаний — не существенно.

Это значит, что каждый случай выбора пары корзин считается столько раз, сколькими способами можно поменять местами шары в уже отмеченных корзинах. В данном случае таких перемен мест (перестановок) всего две.

Итак, выбрать две корзины из четырех можно шестью способами.

Пример 3. Сколькими способами можно выбрать из четырех пронумерованных корзин три.

Решим сначала задачу о размещениях. По нашей формуле имеется 24 способа размещения трех шаров в четырех корзинах: каждое из приведенных на рис. 4.1 размещений двух шаров может быть дополнено двумя разными размещениями третьего шара в любую из оставшихся пустыми корзин.

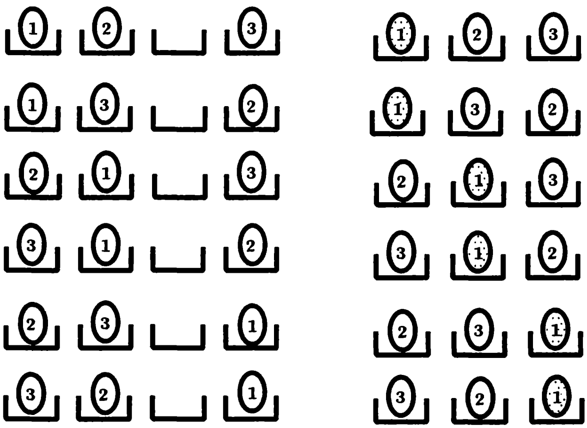


Рис. 4.2

Рассмотрим теперь, сколько раз мы считаем один и тот же выбор корзин, различая размещения разных шаров по одним и тем же корзинам. Рассмотрим, например, размещение, изображенное на рис. 4.2 в первой строке слева. Под ним перечислены все возможные размещения в тех же самых корзинах, но с другим порядком шаров. В правой колонке приведены соответствующие перестановки из трех элементов. Понятно, что эти же самые перестановки будут соответствовать размещениям трех шаров в любых фиксированных трех корзинах. Это значит, что каждый выбор трех корзин из четырех считается шесть раз в формуле числа размещений. Таким образом, три корзины из четырех можно выбрать $\frac{24}{6} = 4$ способами.

Замечание 1′. Отметим интересное равенство: точно так же четырем равно количество возможных выборов из четырех корзин по одной. Это очень важное совпадение объясняется тем, что каждый выбор трех корзин оставляет одну корзину "невыбранной" и, наоборот, каждому выбору одной корзины можно поставить в соответствие выбор дополнительных к ней трех. Следовательно, сколькими способами можно выбрать одну корзину из четырех, столькими способами можно выбрать и три из четырех.

Теперь мы можем решить вопрос о выборе из совокупности в общем случае. Для того чтобы подсчитать, сколькими способами можно вы-

брать k корзин из различимых n, надо сначала вычислить количество размещений k различимых шаров в n корзинах и полученное число поделить на количество перестановок различимых k шаров, или, что то же самое, на количество размещений k шаров в k корзинах. Результат, который называется *числом сочетаний из n элементов по k* и обозначается C_n^k, запишем сначала в следующем виде:

$$C_n^k = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)}{k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 2 \cdot 1}.$$

Эту формулу проще запомнить, если домножить и числитель, и знаменатель на $(n-k)!$, тогда в числителе окажется $n!$ и формула примет вид

$$C_n^k = \\ = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1) \times (n-k) \cdot (n-k-1) \cdot (n-k-2) \cdot \ldots \cdot 2 \cdot 1}{k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 2 \cdot 1 \times (n-k) \cdot (n-k-1) \cdot (n-k-2) \cdot \ldots \cdot 2 \cdot 1} = \\ = \frac{n!}{k! \cdot (n-k)!}.$$

Замечание 1. Из последнего варианта формулы сразу понятно, что $C_n^k = C_n^{n-k}$, поскольку если $a = n-k$, то $n-a = k$ и знаменатели в обеих формулах совпадут. На уровне нашей подразумеваемой интерпретации в терминах выбора корзин этому равенству соответствует следующий аргумент, повторяющий приведенный в предыдущем замечании: если мы выбрали k корзин из n, то тем самым мы выбрали и дополнительные, оставшиеся $n-k$ корзин. Всякому выбору k корзин соответствует единственный выбор $n-k$ корзин.

4.3. Бином Ньютона

Рассмотрим бином $(p+q)^n$. Что получится, если раскрыть скобки? Начнем с выражения $(p+q)^4$. Для того чтобы не потерять ни одного слагаемого, будем использовать так называемое двоичное дерево (рис. 4.3). Каждая ветка дерева заканчивается произведением p и q, которое получается по следующему правилу: если на первом шаге от корня дерева выбирается ветка p, то p становится первым членом произведения, если q, то первым членом произведения становится q. На втором шаге аналогично, в зависимости от выбора ветки, добавляется

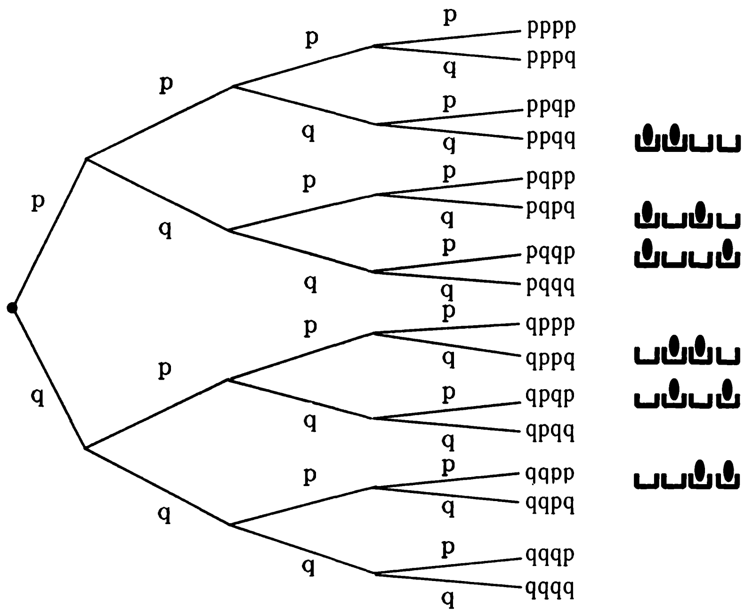


Рис. 4.3

второй сомножитель p или q. Так же осуществляется выбор на третьем и четвертом шагах.

Легко видеть, что общее количество концевых вершин $16 = 2^4$ — два в степени, равной степени бинома. Наш следующий вопрос: сколько слагаемых из этих шестнадцати после перегруппировки сомножителей окажутся равными p^2q^2? Рис. 4.3 показывает, что мы можем дать ответ на этот и подобные вопросы без непосредственного счета — достаточно установить соответствие между каждым из произведений, содержащих p и q во второй степени, и элементами нашей "корзиночной модели": каждому такому произведению соответствует выбор каких-то двух корзин из четырех. На рисунке справа, рядом с произведениями, содержащими квадраты p и q, изображены соответствующие этим произведениям выборы корзинок. Сколько способов выбора пары корзинок из четырех, столько будет членов, содержащих квадраты p и q, в биноме после раскрытия всех скобок.

Совершенно аналогично ведется рассуждение в общем случае: если нам надлежит раскрыть скобки в выражении $(p+q)^n$, то слагаемых,

содержащих $p^k q^{n-k}$ будет столько, сколько существует различных выборов k (или, что то же самое, $n - k$) корзин из n возможных.

Таким образом,

$$(p+q)^n = C_n^0 p^n + C_n^1 p^{n-1} q + C_n^2 p^{n-2} q^2 + \ldots + C_n^{n-1} p q^{n-1} + C_n^n q^n.$$

В этой формуле присутствует знак C_n^0, который словами может быть описан по аналогии со своими соседями: сколькими способами можно выбрать 0 корзин из n. При всей сомнительности такой формулировки из алгебраической формулы понятно, что $C_n^0 = 1$.

По крайней мере, нет никаких соображений, по которым выбрать из n корзин 0 можно каким-то другим, отличным от единицы количеством способов. В подобных ситуациях математики без колебаний принимают соглашение считать C_n^0 равным единице, а заодно равным единице и 0!, который входит в формулу

$$C_n^0 = \frac{n!}{n! \cdot 0!}.$$

4.4. Треугольник Паскаля

На рис. 4.4 изображено несколько видоизмененное двоичное дерево. С его помощью можно находить биномиальные коэффициенты, вообще не используя формулы.

Эта фигура называется треугольником Паскаля.

По двоичному дереву можно передвигаться, стартуя из левой вершины и смещаясь вправо-вверх (выбор p), либо вправо-вниз (выбор q). Сколько существует разных траекторий, приводящих, скажем, в вершину, помеченную номером 6? Заметим, что попасть в эту вершину можно только по такой траектории, в которой выборов p и выборов q одинаковое количество — по два. С другой стороны, любая последовательность, содержащая два p и два q в любом порядке, определяет траекторию, приводящую в данную вершину. Таким образом, разных траекторий, ведущих в вершину, помеченную шестеркой, имеется как раз $C_4^2 = 6$.

Аналогичное рассуждение показывает, что количество траекторий, ведущих в любую вершину, равно какому-то числу сочетаний C_n^k. Нижний индекс равен количеству звеньев в любой траектории, ведущей в данную вершину, а верхний — числу выборов p в этих траекториях.

Проверим, что в каждой вершине нашего дерева записано как раз число траекторий, которые в нее ведут. Начнем от вершин второго

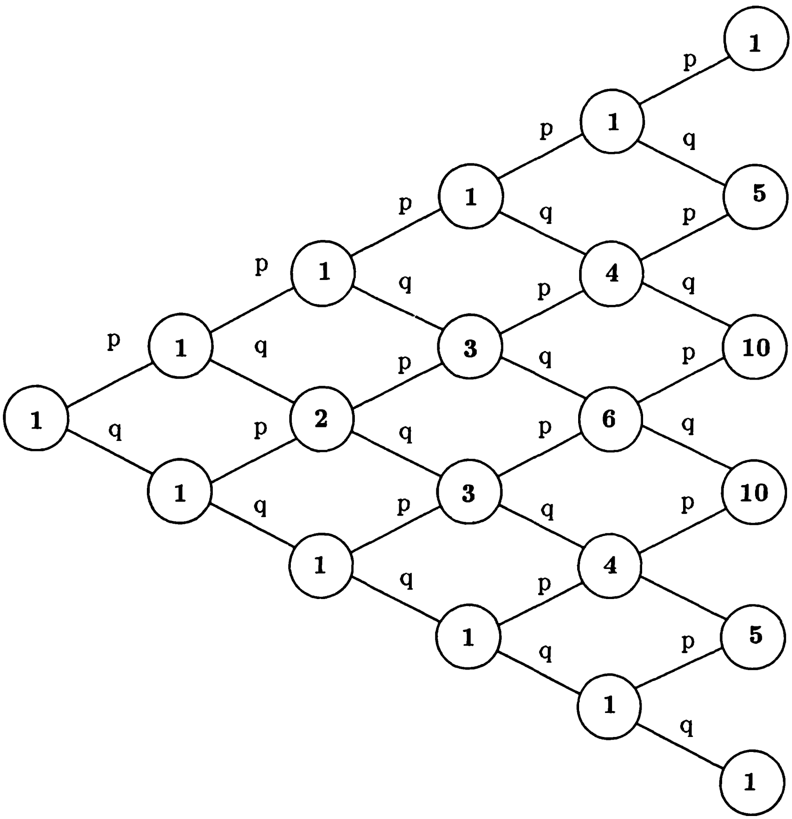


Рис. 4.4

столбца. Очевидно, что в эти вершины из корня дерева действительно ведут по одной траектории. Далее, в верхнюю и нижнюю вершины третьего столбца также ведут по одной траектории, а в среднюю вершину ведет одна траектория из верхней соседней вершины второго столбца и еще одна из нижней соседней вершины второго столбца.

Далее, во вторую сверху вершину четвертого столбца ведут
— одна траектория из верхней соседней вершины третьего столбца
— и каждая (из двух) траектория, попадающая в нижнюю соседнюю вершину третьего столбца может быть продолжена одним звеном p, чтобы окончиться в интересующей нас вершине, справедливо помеченной тройкой.

Это рассуждение можно обобщить. Действительно, если в две соседние вершины некоторого столбца ведут соответственно a и b траекторий, то в находящуюся между ними соседнюю справа вершину приведут $a+b$ разных траекторий, поскольку каждая из ведущих в эти вершины траекторий может быть единственным образом продолжена в нужном направлении.

Поскольку это верно для любой вершины, мы получим правило для нахождения всевозможных биномиальных коэффициентов: в самые верхние и нижние вершины проставляются единицы, а во все остальные, продвигаясь слева направо, надо поставить суммы двух чисел, помещенных в соседние слева вершины.

В заполняемом таким образом бесконечном треугольнике в столбце с номером n (если считать самую левую, корневую вершину нулевым столбцом) на месте k сверху (если считать самую верхнюю вершину столбца нулевой) будет расположено число C_n^k.

4.5. Схема испытаний Бернулли с $p = q = 1/2$

Для того чтобы вычислять простейшие биномиальные вероятности, остается ответить на один простой вопрос: сколько всего траекторий ведут во все вершины столбца с номером n или, что то же самое, какова сумма биномиальных коэффициентов $C_n^0 + C_n^1 + \ldots + C_n^{n-1} + C_n^n$?

Первый способ. Воспользуемся следующим свойством траекторий по двоичному дереву: каждая из ведущих в данную вершину траекторий может быть продолжена двумя способами в вершины следующего столбца. Это значит, что если в вершины n-го столбца ведут в сумме M траекторий, то в следующий столбец ведут ровно $2M$ траекторий. В первый столбец (считая, как договорились, корневую вершину нулевым столбцом) ведут 2 траектории, значит, в n-ю вершину ведет 2^n траекторий. Это означает, что искомая сумма всех биномиальных коэффициентов с одинаковым нижним индексом n также равна 2^n.

Второй способ. Разложим по формуле бинома Ньютона двучлен $(1+x)^n$:

$$(1+x)^n = C_n^0 + C_n^1 x + \ldots + C_n^{n-1} x^{n-1} + C_n^n x^n.$$

Если теперь положить $x = 1$, то равенство примет вид:

$$2^n = C_n^0 + C_n^1 + \ldots + C_n^{n-1} + C_n^n.$$

Мы вывели формулы вероятностей событий “в n испытаниях наблюдается k успехов” в схеме испытаний Бернулли с $p = q = 1/2$. Эта вероятность равна количеству разных последовательностей, содержащих k членов p и $n-k$ членов q (или, что то же самое, количеству траекторий, ведущих в вершину k столбца n), деленной на 2^n:

$$p(1/2, n, k) = \frac{C_n^k}{2^n}.$$

Пока мы использовали буквы p и q только для подсчета траекторий. Поскольку $p = q = 1/2$, то все траектории равновероятны и искомая вероятность действительно равна отношению количества “благоприятных” исходов к общему числу возможных исходов.

Упражнение 4.1. Заполнить столбцы треугольника Паскаля до одиннадцатого включительно. В столбцах с номерами 2, 5, 8, 11 подсчитать суммы, соответствующие средним группам вершин, составляющим третью часть от их общего количества в данном столбце:

$$C_2^1,\ C_5^2 + C_5^3,\ C_8^3 + C_8^4 + C_8^5,\ C_{11}^4 + C_{11}^5 + C_{11}^6 + C_{11}^7.$$

Подсчитать соответствующие вероятности.

Ответ. Вероятности равны соответственно 0,5; 0,625; 0,711; 0,773.

Можно доказать, что доля траекторий, которые оканчиваются в средней трети вершин, стремится к единице при неограниченном росте n.

Больше того, любая фиксированная доля центральных вершин обладает этим свойством. Пусть, заполнив треугольник Паскаля очень далеко вправо, мы расчитываем аналогично предыдущему упражнению вероятности попадания в сотую часть вершин, группирующихся вокруг центральной. Эта вероятность также стремится к единице при увеличении n. Мы докажем это в главе 8.

4.6. Схема испытаний Бернулли с $p \neq q$

Каждая траектория на рис. 4.4 изображает возможный результат последовательности испытаний Бернулли. Последовательности, приводящие в одну вершину, имеют одинаковое количество букв p (так же как и q), поэтому вероятность каждой из них равна $p^k q^{n-k}$, где n и k нумеруют столбец и место вершины в столбце. Любые две такие последовательности представляют несовместимые события, поэтому, для

того чтобы вычислить вероятность попадания в данную вершину, надо сложить вероятности всех последовательностей, приводящих к данному результату. Полученная сумма равна

$$C_n^k p^k q^{n-k}.$$

Глава 5

Случайные величины

5.1. Понятие случайной величины. Закон распределения. Биномиальная случайная величина

Мы уже знакомы с понятиями "испытание", "случайное событие", "вероятность". Теперь мы приступим к рассмотрению чрезвычайно важного случая, который характеризуется следующим обстоятельством: в результате испытания не только происходит событие, но есть еще и возможность наблюдать некоторое число. Причем это число нас интересует даже в большей степени, чем само событие.

Нетрудно видеть, что возможность наблюдать число часто имела место и в тех испытаниях, которые мы уже рассматривали.

Пример 1. Бросая игральный кубик, мы получаем число точек на верхней грани.

Пример 2. Бросая одновременно три монеты, фиксируем число выпадений герба (это может быть 0, 1, 2 или 3).

Если каждому событию подобным образом поставлено в соответствие некоторое число, будем говорить, что задана *случайная величина*. Иными словами, случайная величина — это величина, принимающая в результате испытания то или иное числовое значение, но заранее не известно, какое именно.

Будем обозначать случайные величины большими латинскими буквами X, Y и т.д.

С каждой случайной величиной связано некоторое множество чисел — значений, которые она может принимать. В результате испытания

эти значения могут получаться с различной вероятностью. Правило, устанавливающее связь между возможными значениями и их вероятностями (точнее, речь идет о вероятности события, заключающегося в том, что случайная величина приняла то или иное значение), называется *законом распределения* случайной величины.

> **Замечание 1.** Случайная величина однозначно и полностью определяется своим законом распределения, подобно тому как квадрат определяется длиной стороны. Переводя эту аналогию в плоскость соотношения с реальным миром, заметим, что если квадрат со стороной 10 м может быть моделью дома, бассейна или детской площадки, то случайная величина с данным законом распределения может быть моделью числа посетителей магазина в течение дня, числа выпускаемых станком деталей и т.п.

Вследствие тесной связи между понятиями "случайная величина" и "закон распределения" (или даже просто "распределение"), они часто используются как синонимы.

Перейдем теперь к рассмотрению того, каким образом может быть задан закон распределения случайной величины в случае, когда она принимает лишь конечное число значений.

Итак, пусть случайная величина X может принимать одно из n различных значений

$$x_1, x_2, \ldots, x_n.$$

При этом каждое из этих значений величина X принимает с определенной вероятностью — соответственно

$$p_1, p_2, \ldots, p_n.$$

Иными словами, p_1 — это вероятность случайного события "случайная величина X приняла значение x_1" или, более кратко, $X = x_1$,
p_2 — вероятность случайного события $X = x_2$,
...
p_n — вероятность случайного события $X = x_n$.

Сведем все эти значения в таблицу

X	x_1	x_2	$\ldots$	x_n
	p_1	p_2	$\ldots$	p_n

в первой строке которой указаны значения, принимаемые случайной величиной X, во второй строке — их вероятности. Такая таблица называется *таблицей распределения* случайной величины X. Обычно числа в первой строке таблицы распределения располагают в порядке возрастания.

Замечание 2. Отметим следующее важное обстоятельство. Поскольку в результате испытания величина X наверняка примет одно из этих значений, то сумма несовместимых событий

$$\{X = x_1\} + \{X = x_2\} + \ldots + \{X = x_n\}$$

является достоверным событием, вероятность которого равна 1. Поэтому для таблицы распределения любой случайной величины справедливо равенство

$$p_1 + p_2 + \ldots + p_n = 1.$$

Итак, для того чтобы при решении конкретной задачи заполнить таблицу распределения заданной случайной величины, надо выписать все принимаемые ею значения $x_1, x_2, \ldots, x_n$ и вычислить соответствующие вероятности $p_1, p_2, \ldots, p_n$.

Вернемся к примеру с игральным кубиком. Для упомянутой там случайной величины (обозначим ее через X) вероятности принять любое из шести значений равны между собой. Таблица распределения выглядит так

X	1	2	3	4	5	6
	1/6	1/6	1/6	1/6	1/6	1/6

Для случайной величины из примера с тремя монетами (обозначим ее через Y) построить таблицу распределения несколько сложнее. Вспомним, что в результате одновременного бросания трех монет возможно всего восемь равновероятных исходов: ГГГ, ГГЦ, ГЦГ, ГЦЦ, ЦГГ, ЦГЦ, ЦЦГ, ЦЦЦ. При первом исходе величина Y принимает значение 3; при втором, третьем и пятом — значение 2; при четвертом, шестом и седьмом — значение 1; при восьмом — значение 0. С учетом этого таблица распределения случайной величины Y такова:

Y	0	1	2	3
	1/8	3/8	3/8	1/8

Для более наглядного представления закона распределения часто используется координатная плоскость. По оси абсцисс отмечают значения, принимаемые случайной величиной, на оси ординат — их вероятности. Затем на плоскости (x,p) строят точки $(x_1,p_1),(x_2,p_2),\ldots,(x_n,p_n)$. Для случайной величины Y из примера с тремя монетами это выглядит так, как изображено на рис. 5.1. Если теперь провести от отмеченных точек вертикальные отрезки до пересечения с осью абсцисс, то получится *столбиковая диаграмма* (рис. 5.2). Если же последовательно соединить точки отрезками, получится *полигон* (рис. 5.3).

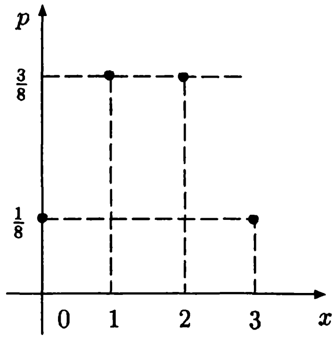


Рис. 5.1

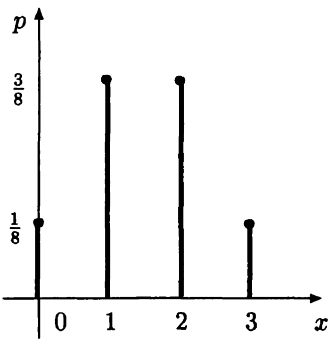


Рис. 5.2

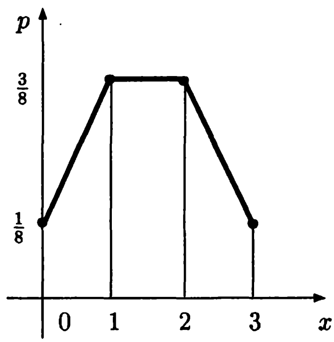


Рис. 5.3

Пример 3 (схема испытаний Бернулли). Испытание повторяется n раз, причем вероятность успеха в одном испытании равна p.

Общее число успехов (в n испытаниях) есть случайная величина, принимающая значения $0,1,2,\ldots,n$. Вероятность того, что эта случайная величина примет значение k, равна

$$P(p,n,k) = C_n^k p^k (1-p)^{n-k}.$$

Такую случайную величину будем называть *биномиальной* и обозначать $B(n,p)$. Она зависит от двух параметров — n и p. Случайная

величина Y из примера 2 на стр. 240 является биномиальной случайной величиной с параметрами 3 (три монеты) и $\frac{1}{2}$ (такова вероятность выпадения герба у одной монеты),

$$Y = B\left(3, \frac{1}{2}\right).$$

5.2. Операции над случайной величиной

Пусть имеется случайная величина X, принимающая в зависимости от результата испытания те или иные случайные значения. Если к каждому из этих значений прибавить одно и то же число, например число 3, то в результате мы получим новые числа — значения случайной величины $X+3$.

Таблица распределения случайной величины $X+3$ строится по таблице распределения случайной величины X следующим образом

X	x_1	...	x_n
	p_1	...	p_n

$\longrightarrow$

$X+3$	x_1+3	...	x_n+3
	p_1	...	p_n

Как видно, вторая строка осталась без изменений, поскольку вероятности событий $(X = x_i)$ и $(X + 3 = x_i + 3)$ равны.

Построение таблицы распределения случайной величины X^2 несколько сложнее.

Рассмотрим конкретный пример

X	−1	0	1	2
	0,2	0,3	0,4	0,1

Таблица для случайной величины $X+3$ строится просто

$X+3$	2	3	4	5
	0,2	0,3	0,4	0,1

Пытаясь действовать аналогичным образом для величины X^2, то есть заменяя все значения x_i числами x_i^2, получаем

	1	0	1	4
	0,2	0,3	0,4	0,1

В первой строке есть совпадающие значения. Поэтому следует объединить их в одно, сложив соответствующие вероятности

X^2	1	0	4
	0,6	0,3	0,1

Расссмотрим еще один пример

Z	−2	−1	0	1	2	3
	0,1	0,2	0,3	0,2	0,1	0,1

Возводя значения случайной величины Z в квадрат, получаем

	4	1	0	1	4	9
	0,1	0,2	0,3	0,2	0,1	0,1

И, наконец, в результате имеем

Z^2	4	1	0	9
	0,2	0,4	0,3	0,1

Таблицу распределения случайной величины $Y = f(X)$ для любой функции f можно построить аналогично. Она строится в два этапа. Сначала вычисляются элементы вспомогательной таблицы

	$f(x_1)$	$f(x_2)$	...	$f(x_n)$
	p_1	p_2	...	p_n

Затем совпадающие значения $f(x_i) = f(x_j)$ для разных чисел x_i и x_j (если такие имеются) объединяются в одно, а соответствующие вероятности складываются.

5.3. Числовые характеристики случайной величины

Как было отмечено ранее, случайная величина полностью определяется своим законом распределения. В некоторых случаях бывает полезно знать некоторые дополнительные числовые характеристики распределения, более того, иногда эти характеристики оказываются даже важнее самого распределния.

Математическое ожидание

Первая важная характеристика — это среднее ожидаемое значение, принимаемое случайной величиной в больших сериях испытаний.

Пусть имеется случайная величина X с заданной таблицей распределения

X	x_1	x_2	...	x_n
	p_1	p_2	...	p_n

Математическое ожидание случайной величины X определяется формулой

$$MX = x_1p_1 + x_2p_2 + \ldots + x_np_n,$$

которая в сокращенной записи выглядит так

$$MX = \sum_{i=1}^{n} x_ip_i.$$

Пояснение 1. Математическое ожидание имеет прозрачный смысл.

Предположим, что проведено m испытаний (m — достаточно большое число), при этом величина X ровно m_1 раз приняла значение x_1, ровно m_2 раза — значение x_2, ..., ровно m_n раз — значение x_n,

$$m_1 + m_2 + \ldots + m_n = m.$$

Найдем среднее арифметическое всех этих m значений. Имеем

$$\frac{\overbrace{x_1 + \ldots + x_1}^{m_1} + \overbrace{x_2 + \ldots + x_2}^{m_2} + \ldots + \overbrace{x_n + \ldots + x_n}^{m_n}}{m} =$$

$$= \frac{x_1m_1 + x_2m_2 + \ldots + x_nm_n}{m} = x_1\frac{m_1}{m} + x_2\frac{m_2}{m} + \ldots + x_n\frac{m_n}{m}.$$

Дробь

$$\frac{m_k}{m}$$

представляет собой относительную частоту появления события $X = x_k$ в m испытаниях (в m испытаниях событие $X = x_k$ произошло m_k раз). При больших значениях m относительная частота примерно равна вероятности события $X = x_k$, то есть p_k, поэтому

$$x_1\frac{m_1}{m} + x_2\frac{m_2}{m} + \ldots + x_n\frac{m_n}{m} \approx x_1p_1 + \ldots + x_np_n = M_X.$$

Таким образом, в серии из большого количестве испытаний среднее арифметическое полученных в этой серии значений случайной величины будет приближаться к ее математическому ожиданию. Этот факт имеет два важных следствия.

Следствие 1. Математическое ожидание случайной величины, распределение которой нам неизвестно, можно оценить средним арифметическим значений в достаточно большой серии ее последовательных испытаний. Больше того, как видно из пояснения 1, чем длинее серия, тем точнее эта оценка. Эта тема будет подробно обсуждаться в следующих разделах.

Следствие 2. В практически интересных случаях серий испытаний можно оценивать наиболее вероятный результат исходя из математического ожидания некоторой случайной величины.

Пример 4. Предлагается следующая азартная игра: Бросают два игральных кубика. Если полученная сумма больше 10, то игрок выигрывает 10 копеек, в противном случае проигрывает 1 копейку. Имеет ли ему смысл играть в эту игру 12 000 000 партий?

Из 36 возможных исходов выпадения двух различимых кубиков в трех случаях выпадает благоприятствующая игроку сумма. Это значит, что 10 копеек он выигрывает с вероятностью $1/12$, а одну копейку проигрывает (скажем так: выигрывает -1 копейку) с вероятностью $11/12$. В таком случае

$$MX = 10 \cdot \frac{1}{12} + (-1) \cdot \frac{11}{12} = -\frac{1}{12}.$$

Проигрывая в среднем $1/12$ копейки за партию, за 12 000 000 партий игрок проиграет около 1 000 000 копеек, или 10 000 рублей.

Пример 5. Предприниматель размышляет над тем, куда лучше вложить деньги — в киоск для торговли мороженым или в палатку для торговли хлебобулочными изделиями.

Вложение средств в киоск с вероятностью $0,5$ обеспечит годовую прибыль 5000 долл., с вероятностью 0,2 — 10 000 долл. и с вероятностью 0,3 — 3000 долл.

Для палатки прогноз таков: 5500 долл. с вероятностью 0,6, 5000 долл. с вероятностью 0,3 и 6500 долл. с вероятностью 0,1.

В каком случае (для киоска или для палатки) математическое ожидание годового дохода больше?

Решение. Для каждого из двух возможных решений годовая прибыль является случайной величиной. Обозначив эти величины X и Y,

построим таблицы распределения

X	3000	5000	10000
	0,3	0,5	0,2

Y	5000	5500	6500
	0,3	0,6	0,1

Найдем математические ожидания

$$MX = 3000 \cdot 0{,}3 + 5000 \cdot 0{,}5 + 10000 \cdot 0{,}2 = 5400 \text{ долл.}$$

$$MY = 5000 \cdot 0{,}3 + 5500 \cdot 0{,}6 + 6500 \cdot 0{,}1 = 5450 \text{ долл.}$$

Получается, что $MY > MX$. Таким образом, математическое ожидание для киоска больше.

Дисперсия и среднеквадратическое (стандартное) отклонение

Итак, математическое ожидание обозначает, какое значение случайная величина принимает “в среднем”. Следующий простейший пример показывает, что случайные величины с равным математическим ожиданием могут существенно различаться по степени близости к нему.

Рассмотрим случайные величины X и Y

X	99	101
	0,5	0,5

Y	0	200
	0,5	0,5

Нетрудно видеть, что $MX = MY = 100$. Но если для величины X отклонение от значения 100 незначительно, то для величины Y оно весьма заметно.

Если выбор между величинами X и Y — это выбор между двумя альтернативными решениями, то X — это более стабильный, предсказуемый результат, а Y — это в большей степени риск.

Показателем этой “непредсказуемости” служит еще одна числовая характеристика случайной величины, называемая *дисперсией*. Обозначение: DX (от англ. dispersion). Покажем, как она вычисляется.

Вычитая из случайной величины X ее математическое ожидание (которое является числом — в предыдущем нашем примере это число 100), получаем новую случайную величину

$$X - MX.$$

Квадрат последней также является случайной величиной

$$(X - MX)^2,$$

математическое ожидание которой и есть дисперсия X

$$DX = M(X - MX)^2.$$

Если величина X задана таблицей

X	x_1	x_2	...	x_n
	p_1	p_2	...	p_n

то дисперсия случайной величины X может быть вычислена по формуле

$$DX = \sum_{i=1}^{n} (x_i - MX)^2 p_i$$

или более просто

$$DX = \sum_{i=1}^{n} x_i^2 p_i - (MX)^2.$$

Пример 6. Вычислим дисперсии случайных величин X и Y, таблицы которых приведены на стр. 248

$$DX = 99^2 \cdot 0{,}5 + 101^2 \cdot 0{,}5 - 100^2 = 1,$$
$$DX = 0^2 \cdot 0{,}5 + 200^2 \cdot 0{,}5 - 100^2 = 10\,000.$$

Пример 7. Для биномиальной случайной величины $X = B(n,p)$ справедливы соотношения

$$MX = np, \qquad DX = np(1 - p).$$

Замечание 3. Математическое ожидание может быть любым числом, а дисперсия всегда неотрицательна.

Случайные величины, моделирующие какие-либо объекты реального мира, обычно имеют размерность. Это означает, что принимаемые ими значения могут измеряться в штуках, метрах, килограммах и т.п. При этом математическое ожидание случайной величины имеет ту же размерность, что и сама случайная величина. Размерность же дисперсии равна квадрату размерности случайной величины. Например, если случайная величина измеряется в рублях, то ее дисперсия — в рублях в квадрате.

Чтобы не иметь дело с такими причудливыми единицами измерения, вводится понятие среднеквадратического (*стандартного*) отклонения. Оно обозначается греческой буквой σ (сигма) и по определению равно квадратному корню из дисперсии

$$\sigma = \sqrt{DX}.$$

Тем самым стандартное отклонение имеет ту же размерность, что и сама случайная величина.

Выше шла речь об операциях над случайными величинами. В случае линейных преобразований случайной величины X (то есть преобразований вида

$$Y = aX + b,$$

где a и b — некоторые числа) математическое ожидание и дисперсию получившейся случайной величины Y можно вычислить, исходя из этих же числовых характеристик величины X. Именно, справедливы следующие формулы:

$$MY = aMX + b,$$
$$DY = a^2 DY.$$

Стандартная случайная величина

Случайная величина, у которой математическое ожидание равно 0, а дисперсия равна 1, называется *стандартной* или *стандартизованной* случайной величиной.

Пусть имеется случайная величина X с математическим ожиданием μ (читается “мю”) и стандартным отклонением σ. Нетрудно показать, что случайная величина

$$Y = \frac{X - \mu}{\sigma}$$

является стандартной случайной величиной.

5.4. Сумма случайных величин

Если в результате испытания принимают свои значения сразу две случайные величины, X_1 и X_2, то можно их рассматривать вместе. В частности, определим их сумму $X_1 + X_2$ следующим образом: если в результате испытания величина X_1 принимает значение x_1, а величина X_2 — значение x_2, то случайная величина $X_1 + X_2$ принимает значение $x_1 + x_2$.

Аналогично определяется сумма n случайных величин.

Пример 8. Студент сдает в сессию три экзамена. Оценка по i-му экзамену ($i = 1, 2, 3$) — случайная величина X_i. Тогда общая сумма оценок за все три экзамена — случайная величина $X_1 + X_2 + X_3$.

Приведем пример построения закона распределения суммы двух случайных величин.

Пример 9. Пусть случайные величины X_1 и X_2 задаются таблицами распределения следующего вида:

X_1	-1	0	1
	0,2	0,5	0,3

X_2	-1	1
	0,4	0,6

Пусть также у нас есть основания считать, что случайные величины физически не зависят одна от другой. Так как X_1 принимает три различных значения, а X_2 — два, то для суммы $X_1 + X_2$ получаем шесть возможностей. Выпишем их и вычислим попутно вероятности, используя независимость событий "случайная величина X_1 принимает i-е значение и "случайная величина X_2 принимает j-е значение" и перемножая соответствующие вероятности:

$$p(X_1 = -1)\, p(X_2 = -1) = 0{,}2 \cdot 0{,}4 = 0{,}08,$$

при этом $\quad X_1 + X_2 = -2;$

$$p(X_1 = -1)\, p(X_2 = 1) = 0{,}2 \cdot 0{,}6 = 0{,}12,$$

при этом $\quad X_1 + X_2 = 0;$

$$p(X_1 = 0)\, p(X_2 = -1) = 0{,}5 \cdot 0{,}4 = 0{,}2,$$

при этом $\quad X_1 + X_2 = -1;$

$$p(X_1 = 0)\, p(X_2 = 1) = 0{,}5 \cdot 0{,}6 = 0{,}3,$$

при этом $\quad X_1 + X_2 = 1;$

$$p(X_1 = 1)\, p(X_2 = -1) = 0{,}3 \cdot 0{,}4 = 0{,}12,$$

при этом $\quad X_1 + X_2 = 0;$

$$p(X_1 = 1)\, p(X_2 = 1) = 0{,}3 \cdot 0{,}6 = 0{,}18,$$

при этом $\quad X_1 + X_2 = 2.$

Таким образом, для величины X_1+X_2 получаем следующую таблицу распределения:

X_1+X_2	-2	-1	0	1	2
	0,08	0,2	0,24	0,3	0,18

В случае суммирования n дискретных случайных величин, где $n > 2$, таблица распределения строится аналогичным образом. Как в дискретном, так и в непрерывном случаях для математического ожидания суммы $X_1+X_2+\ldots+X_n$ справедлива следующая простая формула:

$$M\left(X_1+\ldots+X_n\right)=MX_1+\ldots+MX_n$$

или, применяя знак сокращенного суммирования,

$$M\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}MX_i.$$

Эта формула для математического ожидания справедлива для любых случайных величин. Аналогичная формула для дисперсии справедлива не всегда, а только в случае *независимых* случайных величин.

Случайные величины $X_1,X_2,\ldots,X_n$ называются независимыми, если закон распределения каждой из них не зависит от того, какие значения приняли другие величины.

Для дисперсии суммы независимых случайных величин справедливо соотношение

$$D\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}DX_i.$$

Упражнение 5.1. Равны ли дисперсии случайных величин $2X$ и $X+X$? *Ответ.* Запись $X+X$ не вполне корректна, обычно мы писали X_1+X_2 с оговоркой, что эти случайные величины распределены так же, как X. После этого ясно, что дисперсия X_1+X_2 равна $2D$, а дисперсия $2X$ равна $4D$.

Замечание 4. Отметим, что величины $X_1,X_2,\ldots,X_n$ могут иметь один и тот же закон распределения. В этом случае вместо термина “независимые случайные величины” употребляют термин “независимые наблюдения (испытания)”.

5.5. Случайные величины с бесконечным числом значений

Выше мы рассматривали случайные величины, принимающие конечное число значений. Однако часто возникает необходимость рассмотрения случайных величин, число возможных значений которых бесконечно.

Приведем два простых примера.

Пример 10. Испытание состоит в бросании монеты до первого выпадения герба. Случайная величина Y — количество бросаний — может принимать значения 1, 2, 3 и так далее, т.е. бесконечное число значений.

Пример 11. Испытание состоит в том, что на отрезке $[0; 1]$ числовой оси случайным образом отмечается точка (изначально все точки отрезка "равноправны", то есть шансы каждой точки быть отмеченной такие же, как у любой другой). Случайная величина X — число отрезка $[0; 1]$, соответствующее отмеченной точке — может принимать любые значения из отрезка $[0; 1]$.

Натуральных чисел бесконечно много, однако на числовой прямой они расположены изолировано друг от друга (дискретно). Отмеченное свойство объединяет величину Y из примера 10 со случайными величинами, принимающими конечное число значений. Все эти величины называются *дискретными* случайными величинами.

Для случайной величины Y из первого примера можно построить "бесконечную таблицу распределения" следующего вида

Y	1	2	3	...	k	...
	1/2	1/4	1/8	...	$(1/2)^k$	...

В отличие от дискретных случайных величин, случайная величина X из второго примера является *непрерывной* случайной величиной — точки отрезка $[0; 1]$ нельзя одну за другой выделить и записать в таблицу (хотя бы и бесконечную).

Существуют еще случайные величины *смешанного типа.*

Далее мы будем рассматривать только случайные величины, принимающие конечное число значений, и непрерывные случайные величины.

5.6. Непрерывные случайные величины

Под *непрерывной случайной величиной* мы будем понимать случайную величину, принимающую значения на прямой, луче (полупрямой) или отрезке. Описание закона распределения в непрерывном случае существенно сложнее, чем в дискретном.

Главное отличие в задачах вычисления вероятностей для дискретного и непрерывного случаев состоит в следующем. В дискретном случае ищется вероятность событий типа $X = c$ (случайная величина принимает определенное значение). В непрерывном случае вероятности такого типа равны нулю, поэтому интерес представляют вероятности событий типа $a \leqslant X \leqslant b$ (случайная величина принимает значения из некоторого отрезка). При этом

$$\begin{aligned} &p(a \leqslant X \leqslant b) = p(a < X \leqslant b) = \\ &\qquad = p(a \leqslant X < b) = p(a < X < b), \\ &p(X \geqslant c) = p(X > c), \\ &p(X \leqslant c) = p(X < c). \end{aligned}$$

Для случайной величины X, с равной вероятностью принимающей любое значение из отрезка $[0; 1]$, естественно считать, что вероятность попадания в отрезок $[a; b]$ равна длине этого отрезка. Например,

$$\begin{aligned} &p(0 \leqslant X \leqslant 0{,}5) = 0{,}5, \\ &p(0{,}5 \leqslant X \leqslant 0{,}7) = 0{,}2, \\ &p(0 \leqslant X \leqslant 1) = 1, \\ &p(X \geqslant 0{,}8) = p(X \leqslant 0{,}2) = 0{,}2. \end{aligned}$$

Рассмотрим функцию

$$f(x) = \begin{cases} 0, & \text{при } x < 0; \\ 1, & \text{при } 0 \leqslant x \leqslant 1; \\ 0, & \text{при } x > 1 \end{cases}$$

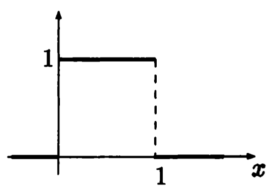


Рис. 5.4

(рис. 5.4). Ее связывает со случайной величиной X следующее обстоятельство: вероятность события $a \leqslant X \leqslant b$ равна площади фигуры, ограниченной прямыми $y = 0$, $x = a$, $x = b$ и графиком функции $y = f(x)$. Иными словами, справедливо равенство

$$p(a \leqslant X \leqslant b) = \int_a^b f(x)dx,$$

для любых a и b, $a \leqslant b$.

Таким образом, функция $f(x)$ позволяет вычислять вероятности, связанные со случайной величиной X, т.е. по сути задает закон распределения случайной величины X.

Посмотрим теперь на ситуацию с более общей точки зрения. Пусть имеется случайная величина X и неотрицательная функция $f(x)$, такая, что для любых чисел a и b, $a \leqslant b$, выполняется равенство

$$p(a \leqslant X \leqslant b) = \int_a^b f(x)dx,$$

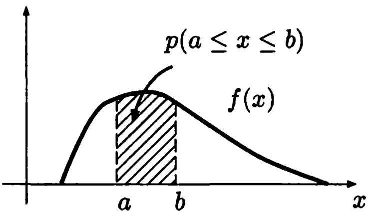


Рис. 5.5

(рис. 5.5). В этом случае говорят, что случайная величина X имеет *плотность распределения* $f(x)$. Записывается это следующим образом:

$$X \sim f(x).$$

Замечание 5. Интеграл от плотности распределения по всей области возможных значений случайной величины равен 1.

Замечание 6. Описанная ситуация допускает следующую аналогию из механики. Пусть имеется сплошной стержень массой 1 кг. Масса распределена по стержню с различной, вообще говоря, плотностью. Если мы хотим найти, сколько весит некоторый отрезок стержня, надо взять интеграл от плотности на этом отрезке.

Пример 12. Пусть случайная величина X задана плотностью распределения

$$f(x) = \begin{cases} 0, & \text{при } x < 0; \\ x/2, & \text{при } 0 \leqslant x \leqslant 2; \\ 0, & \text{при } x > 2 \end{cases}$$

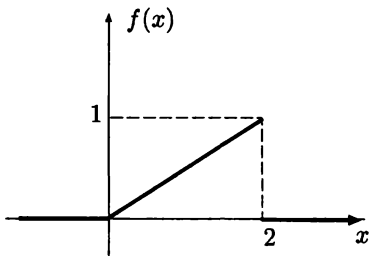


Рис. 5.6

(рис. 5.6). Для нахождения вероятностей требуется вычислять соответствующие интегралы. Например,

$$P(0 \leqslant X \leqslant 1) = \int\limits_0^1 \frac{x}{2} dx = \frac{1}{4} x^2 \Big|_0^1 = \frac{1}{4} - 0 = \frac{1}{4}.$$

Важно отметить, что

$$P(0 \leqslant X \leqslant 2) = 1,$$

поскольку все возможные значения величины X лежат в пределах от 0 до 2.

Пример 13. *Показательное (экспоненциальное) распределение* задается плотностью вида

$$f(x) = \begin{cases} 0, \text{ при } x < 0; \\ \lambda e^{-\lambda x}, \text{ при } x \geqslant 0 \end{cases}$$

(рис. 5.7). Здесь λ — некоторое положительное число. Показательное распределение применяется в *теории массового обслуживания*.

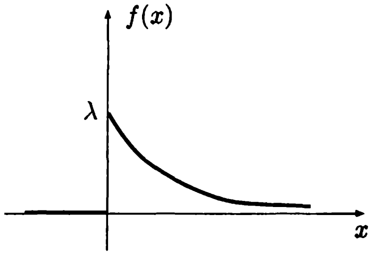


Рис. 5.7

Для непрерывных распределений, как и для дискретных, можно рассматривать числовые характеристики: математическое ожидание, дисперсию и другие. Они вычисляются при помощи плотности распределения. Если случайная величина X имеет плотность распределения $p(x)$, то

$$MX = \int\limits_{-\infty}^{\infty} x\, p(x)\, dx.$$

После этого можно вычислить и дисперсию:

$$DX = \int\limits_{-\infty}^{\infty} (x - MX)^2\, p(x)\, dx.$$

При том, что на первый взгляд формулы для математического ожидания и дисперсии в дискретном и непрерывном случаях кажутся совершенно разными, на самом деле они очень похожи. В следующей главе мы разберем эту аналогию.

Глава 6

О формулах для непрерывных и дискретных случайных величин

В этой главе мы разберем аналогии между непрерывными и дискретными случайными величинами. Вспомним, что площади криволинейных фигур, к которым относятся и интегралы от плотности распределения случайной величины, могут быть вычислены двумя способами — через нахождение первообразной и через предел приближенных значений площади, когда криволинейные трапеции заменяются на мало отличающиеся от них прямоугольники.

Пример такого приближенного значения интеграла от функции показан на рис. 6.1. Жирная линия изображает функцию

$$f(x) = \begin{cases} 0 & \text{при} \quad x < 1; \\ 1 + x & \text{при} \quad -1 \geqslant x < 0; \\ 1 - x & \text{при} \quad 0 \geqslant x < 1; \\ 0 & \text{при} \quad 1 \geqslant x. \end{cases}$$

Отрезок $[-1; 1]$ поделен на десять частей, и на каждом из десяти маленьких отрезков построен прямоугольник с высотой, равной среднему значению функции $f(x)$ на этом маленьком отрезке. В данном случае это среднее значение равно значению функции в середине маленького

отрезка. Таким образом, высоты прямоугольников равны (слева направо) 0,1, 0,3, 0,5, 0,7, 0,9, 0,9, 0,7, 0,5, 0,3, 0,1. Сумма площадей прямоугольников равна (поскольку все они имеют основания, равные 0,2) сумме их высот, умноженной на основание. Сложив высоты, получаем 5, и суммарная площадь равна единице.

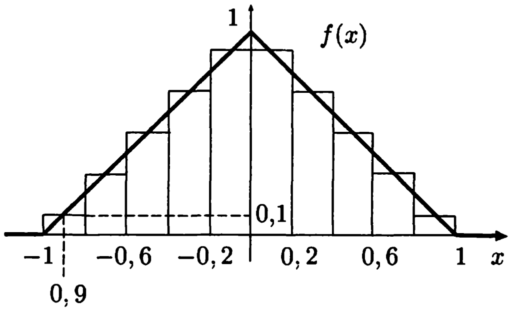


Рис. 6.1

То же самое значение дает и интеграл от $f(x)$:

$$MX = \int_{-\infty}^{\infty} f(x)\,dx = \int_{-1}^{1} f(x)\,dx = 1.$$

Полное совпадение результатов объясняется простым видом функции; если бы ее график был “более криволинеен”, то вычисление с помощью прямоугольников дало бы лишь приближенный результат.

Если случайная величина X имеет плотность распределения $f(x)$, то вероятность $p(-1 < x < -0{,}8)$ равна (в общем случае это равенство приближенное) площади соответствующего прямоугольника, т.е. $0{,}2 \cdot 0{,}1 = 0{,}02$. Аналогично вероятности попадания случайной величины в остальные девять маленьких отрезков равны соответственно (слева направо) 0,06, 0,1, 0,14, 0,18, 0,18, 0,14, 0,1, 0,06, 0,02.

Зададим дискретную случайную величину Y, используя эти площади как вероятности, а в качестве соответствующих этим вероятностям значений возьмем середины отрезочков. Случайная величина Y задается таблицей

$-0,9$	$-0,7$	$-0,5$	$-0,3$	$-0,1$	$0,1$	$0,3$	$0,5$	$0,7$	$0,9$
$0,02$	$0,06$	$0,1$	$0,14$	$0,18$	$0,18$	$0,14$	$0,1$	$0,06$	$0,02$

Поскольку сумма чисел в нижней строке равна единице, наша случайная величина задана корректно. Вычислим теперь математические ожидания величин X и Y.

$$MX = \int_{-\infty}^{\infty} x\,f(x)dx = \int_{-1}^{0} x(x+1)\,dx + \int_{0}^{1} x(1-x)\,dx =$$

$$= \int_{-1}^{0} (x^2+x)\,dx + \int_{0}^{1} (x-x^2)\,dx =$$

$$= \frac{1}{3}x^3\Big|_{-1}^{0} + \frac{1}{2}x^2\Big|_{-1}^{0} + \frac{1}{2}x^2\Big|_{0}^{1} + \frac{1}{3}(-x^3)\Big|_{0}^{1} =$$

$$= \frac{1}{3} - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = 0$$

По нашему построению дискретная случайная величина Y приближенно равна непрерывной случайной величине X: там, где X принимает значения из малого отрезка, Y с той же вероятностью попадает ровно в середину этого отрезка (например, для отрезка $[0,2;\ 0,4]$) это 0,3).

Нет ничего удивительного, что обе случайные величины имеют похожие математические ожидания: для случайной величины Y это

$$MY = 0{,}02\cdot(-0{,}9)+0{,}06\cdot(-0{,}7)+0{,}1\cdot(-0{,}5)+0{,}14\cdot(-0{,}3)+0{,}18\cdot(-0{,}1)+$$

$$+0{,}18\cdot 0{,}1+0{,}14\cdot 0{,}3+0{,}1\cdot 0{,}5+0{,}06\cdot 0{,}7+0{,}02\cdot 0{,}9=0$$

(в общем случае равенство было бы приближенным).

В силу симметрии обоих распределений вычисления можно было и не проводить — результат заведомо должен был быть нулевым. Однако эти вычисления показывают аналогию между дискретным и непрерывным распределением.

Мы провели "дискретизацию" непрерывной случайной величины, поделив ее область значений на отрезки и присвоив дискретной величине значения, совпадающие с серединами этих отрезков, а соответствующие этим значениям вероятности положили равными вероятности попасть в данный отрезок для непрерывной величины.

Мы замечаем далее, что в маленьком отрезке мало меняется как значение функции $f(x)$, так и значение самой переменной x. Возьмем интеграл по этому малому отрезку, который составляет часть интеграла, вычисляющего математическое ожидание

$$\int_{0,2}^{0,4} x\,f(x)\,dx,$$

и заменим под интегралом множитель x, который меняется от 0,2 до 0,4 на мало от них отличающееся среднее значение $0,3$ и вынесем эту константу за знак интеграла. Получим приблизительно равный интеграл:

$$0,3\int_{0,2}^{0,4} f(x)\,dx.$$

Если обозначить $P_Y(0,3)$ вероятность того, что дискретная случайная величина Y примет значение $0,3$, то по нашему построению Y эта вероятность как раз и есть вероятность того, что X попадет в отрезок $[0,2;\,0,4]$, т.е.

$$\int_{0,2}^{0,4} f(x)\,dx.$$

Таким образом,

$$0,3\int_{0,2}^{0,4} f(x)\,dx.$$

представляет собой одно из десяти слагаемых

$$0,3\cdot P_Y(0,3),$$

входящих в формулу вычисления математического ожидания Y, и одновременно приблизительно равен

$$\int_{0,2}^{0,4} x\,f(x)\,dx,$$

представляющего собой одну из десяти составляющих интеграла в формуле вычисления математического ожидания для непрерывной случайной величины X.

Итак, все десять слагаемых двух формул приблизительно равны.

Точно такую же процедуру мы можем проделать и с формулой дисперсии. В формуле

$$\int_{0,2}^{0,4} (x - MX)^2\,f(x)\,dx$$

можно заменить x на $0,3$, в результате чего получится

$$(0,3 - MX)^2 \int_{0,2}^{0,4} f(x)\,dx = (0,3 - MX)P_Y(0,3).$$

Мы помним, что в нашем случае $MX = MY = 0$, в общем же случае мы имеем приблизительное равенство $MX \approx MY$, поэтому

$$(0,3 - MX)P_Y(0,3) \approx (0,3 - MY)P_Y(0,3).$$

Таким образом,

$$\int_{0,2}^{0,4} (x - MX)^2 f(x)\,dx \approx (0,3 - MY)^2 P_Y(0,3).$$

Подобным образом можно доказать приближенное равенство составляющих в формулах дисперсии для любой непрерывной случайной величины и ее "дискретизации", произведенной с помощью "разрезания" области ее значений на малые отрезки. Чем мельче будут взятые отрезки, тем большая точность будет достигнута в приближенных равенствах.

Глава 7

Случайные величины (продолжение)

7.1. Нормальное распределение

Наиболее часто применяется для анализа реальных ситуаций так называемое *нормальное* (или *гауссово*) распределение. Это распределение зависит от двух параметров μ и σ и задается плотностью вида

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(рис. 7.1). Случайную величину, распределенную нормально с параметрами μ и σ, будем обозначать $N(\mu, \sigma)$. Параметры μ и σ имеют вполне ясный смысл: это соответственно математическое ожидание и стандартное отклонение. Зависимость нормального распределения от параметров мы обсудим несколько позже, сначала же рассмотрим важный частный случай.

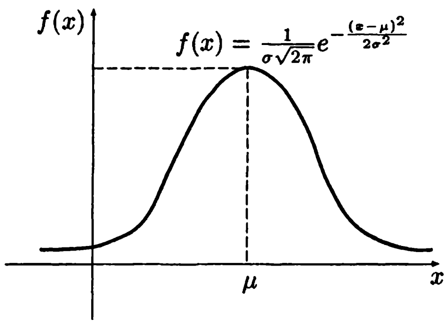


Рис. 7.1

При $\mu = 0$ и $\sigma = 1$ получается *стандартное нормальное распределение* $N(0, 1)$ (напомним, что стандартной называется случайная

величина с нулевым математическим ожиданием и единичной дисперсией). Его плотность будем обозначать особым образом, $\varphi(x)$, а саму случайную величину — U. Ясно, что

$$U \sim \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Функция $\varphi(x)$ четная, то есть $\varphi(-x) = \varphi(x)$, и, следовательно, график симметричен относительно оси оси ординат (рис. 7.2). Поэтому для исследования функции $\varphi(x)$ достаточно рассмотреть ее для значений $x \geqslant 0$. В точке $x = 0$ функция $\varphi(x)$ имеет максимум

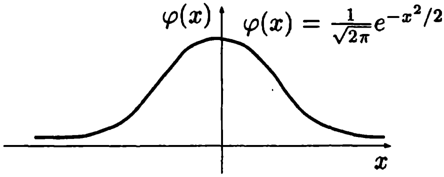


Рис. 7.2

$$\varphi(0) = \frac{1}{\sqrt{2\pi}},$$

а с увеличением аргумента x убывает, причем это убывание происходит довольно быстро.

x	0,0	0,5	1,0	1,5	2,0	2,5	3,0
$\varphi(x)$	0,399	0,352	0,242	0,130	0,054	0,018	0,001

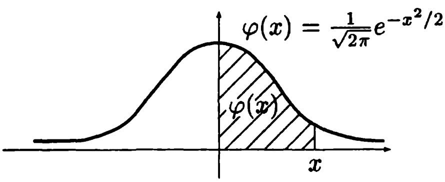


Рис. 7.3

Перейдем к рассмотрению основной задачи — вычислению вероятностей типа

$$p(a \leqslant U \leqslant b).$$

С этой целью определим для неотрицательных чисел x функцию $\Phi(x)$. Имеем

$$\Phi(x) = \int_0^x \varphi(t)dt.$$

Площадь фигуры, заштрихованной на рис. 7.3, равна $\Phi(x)$. Иначе говоря, функция $\Phi(x)$ — первообразная функции $\varphi(x)$.

Для вычислений, связанных с функцией $\Phi(x)$, обычно пользуются таблицами различной степени подробности. Мы будем пользоваться следующей таблицей (для удобства она повторена в Приложении)

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
$0,1$	$0,040$	$1,1$	$0,364$	$2,1$	$0,482$
$0,2$	$0,079$	$1,2$	$0,385$	$2,2$	$0,486$
$0,3$	$0,118$	$1,3$	$0,403$	$2,3$	$0,489$
$0,4$	$0,155$	$1,4$	$0,419$	$2,4$	$0,492$
$0,5$	$0,192$	$1,5$	$0,433$	$2,5$	$0,494$
$0,6$	$0,226$	$1,6$	$0,445$	$2,6$	$0,495$
$0,7$	$0,258$	$1,7$	$0,455$	$2,7$	$0,497$
$0,8$	$0,288$	$1,8$	$0,464$	$2,8$	$0,497$
$0,9$	$0,316$	$1,9$	$0,471$	$2,9$	$0,498$
$1,0$	$0,341$	$2,0$	$0,477$	$3,0$	$0,499$

Можно считать, что при $x > 3$ имеем $\Phi(x) \approx 0,5$.

При помощи этой таблицы можно сразу найти, например, следующие вероятности:

$$p(0 \leqslant U \leqslant 0{,}5) = \Phi(0{,}5) = 0{,}192,$$

$$p(0 \leqslant U \leqslant 0{,}3) = \Phi(0{,}3) = 0{,}118.$$

Из симметричности графика плотности распределения (рис. 7.2), легко видеть, что

$$p(U \geqslant 0) = p(U \leqslant 0) = 0{,}5.$$

Поэтому

$$p(U \geqslant 0,7) = 0,5 - p(0 \leqslant U \leqslant 0,7) = 0,5 - \Phi(0,7) =$$

$$= 0{,}5 - 0{,}258 = 0{,}242,$$

$$p(U \geqslant 0{,}4) = 0{,}5 - p(0 \leqslant U \leqslant 0{,}4) = 0{,}5 - \Phi(0{,}4) =$$

$$= 0{,}5 - 0{,}155 = 0{,}345$$

(рис. 7.4).

Поскольку график функции $\varphi(x)$ симметричен, справедливы следующие соотношения:

$$p(-0,4 \leqslant U \leqslant 0) = p(0 \leqslant U \leqslant 0,4) = \Phi(0,4) = 0,155,$$

$$p(-0,8 \leqslant U \leqslant 0) = p(0 \leqslant U \leqslant 0,8) = \Phi(0,8) = 0,288$$

(рис. 7.5).

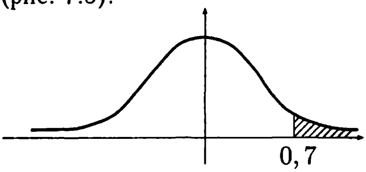


Рис. 7.4

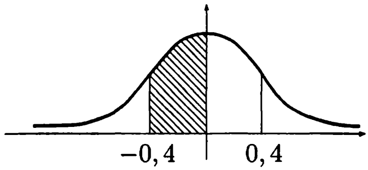


Рис. 7.5

Наконец, вот еще три случая:

$$p(0,1 \leqslant U \leqslant 0,3) = p(-0,3 \leqslant U \leqslant -0,1) =$$

$$= \Phi(0,3) - \Phi(0,1) = 0,118 - 0,040 = 0,078,$$

$$p(-0,1 \leqslant U \leqslant 0,3) = \Phi(0,3) + \Phi(0,1) =$$

$$= 0,118 + 0,040 = 0,158.$$

(рис. 7.6, 7.7, 7.8).

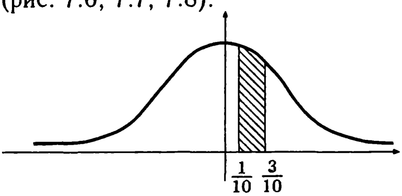


Рис. 7.6

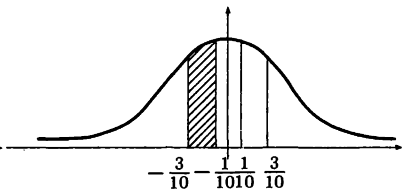


Рис. 7.7

Замечание 1. Во всех этих формулах знаки “$\leqslant$” и “$\geqslant$” можно заменить на “$<$” и “$>$” соответственно.

Перейдем теперь к рассмотрению нормального распределения общего вида. Напомним, что нормальное распределение с математическим ожиданием μ и стандартным отклонением σ обозначается через $N(\mu, \sigma)$.

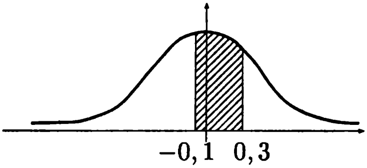


Рис. 7.8

На рис. 7.9, 7.10 показано, как график плотности нормального распределения зависит от параметров μ и σ. Чем больше μ, тем “правее” расположен график (при одинаковых σ). Чем больше σ, тем график более “пологий” (при одинаковых μ).

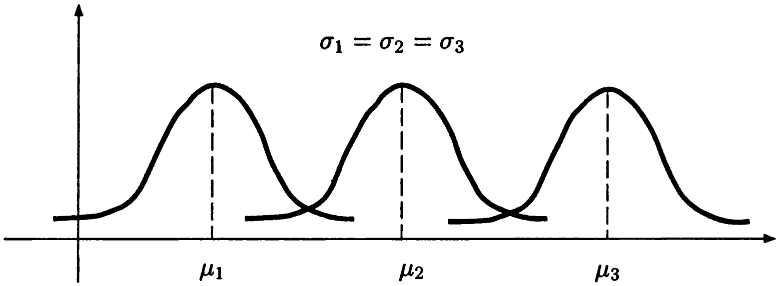


Рис. 7.9

Пусть имеется случайная величина $X = N(\mu, \sigma)$. Вычисление вероятностей типа $p(a \leqslant X \leqslant b)$ сводится к вычислению аналогичных вероятностей для стандартной величины $U = N(0,1)$. Оказывается, что линейные операции над нормальной случайной величиной приводят опять к нормальной случайной величине (с другими числовыми характеристиками). Для нас сейчас важно то, что, *вычитая из нормальной случайной величины ее математическое ожидание и деля на стандартное отклонение, получаем в результате стандартную нормальную величину U*

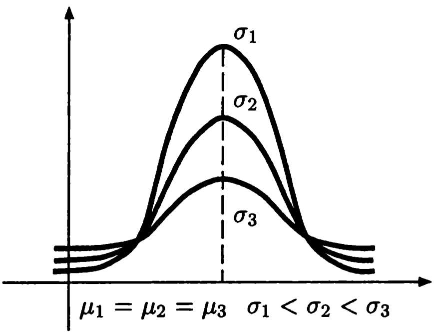


Рис. 7.10

$$\frac{X-\mu}{\sigma} = U.$$

Поскольку неравенства

$$a \leqslant X \leqslant b \quad \text{и} \quad \frac{a-\mu}{\sigma} \leqslant \frac{X-\mu}{\sigma} \leqslant \frac{b-\mu}{\sigma}$$

выполняются или не выполняются всегда одновременно, то

$$p(a \leqslant X \leqslant b) = p\left(\frac{a-\mu}{\sigma} \leqslant \frac{X-\mu}{\sigma} \leqslant \frac{b-\mu}{\sigma}\right) =$$

$$= p\left(\frac{a-\mu}{\sigma} \leqslant U \leqslant \frac{b-\mu}{\sigma}\right).$$

Аналогично

$$p(X \leqslant c) = p\left(\frac{X-\mu}{\sigma} \leqslant \frac{c-\mu}{\sigma}\right) = p\left(U \leqslant \frac{c-\mu}{\sigma}\right),$$

$$p(X \geqslant c) = p\left(\frac{X-\mu}{\sigma} \geqslant \frac{c-\mu}{\sigma}\right) = p\left(U \geqslant \frac{c-\mu}{\sigma}\right).$$

Пример 1. Вес пачек с рисом, расфасованных фирмой “Новый рис”, является нормально распределенной случайной величиной с математическим ожиданием 1 кг и стандартным отклонением 10 г. Какой процент пачек имеет вес:

а) в промежутке от 990 г до 1020 г;

Вес пачки риса является нормально распределенной величиной $X = N(1000, 10)$. Поэтому

$$p\,(990 < X < 1020) = p\left(\frac{990-1000}{10} < U < \frac{1020-1000}{10}\right) =$$

$$= p\,(-1 < U < 2) = \Phi(1) + \Phi(2) = 0,341 + 0,477 = 0,818;$$

б) более 1020 г?

$$p\,(X > 1020) = p\left(U > \frac{1020-1000}{10}\right) =$$

$$= p\,(U > 2) = 0,5 - \Phi(2) = 0,5 - 0,477 = 0,023.$$

в) менее 990 г?

$$p\,(X < 990) = p\left(U < \frac{990-1000}{10}\right) =$$

$$= p\,(U < -1) = 0,5 - \Phi(1) = 0,5 - 0,341 = 0,159;$$

Наряду с вычисленем вероятностей, связанных с нормальным распределением, можно решать задачи и другого типа.

Пример 2. Найти такое число C, что

$$p(0 < U < C) = 0,2.$$

Решение. Имеем

$$p(0 < U < C) = \Phi(C).$$

По условию

$$\Phi(C) = 0,2.$$

Обратимся к таблице значений функции Ф. Среди ее значений нет числа 0,2. Поэтому ищем ближайшее к нему. Это число 0,192, соответствующее значению $C = 0,5$.

Ответ: $C = 0,5$.

Пример 3. Найти такое число C, что

$$p(U < C) = 0,1.$$

Решение. Для любого положительного числа С выполняется неравенство

$$p(U < C) > 0,5.$$

Поэтому искомое значение C — число отрицательное. С учетом этого получаем, что

$$p(U < C) = 0,5 - \Phi(-C)$$

(в качестве аргумента функции Ф указано положительное число $-C$). Далее

$$0,5 - \Phi(-C) = 0,1;$$

$$\Phi(-C) = 0,4.$$

Среди значений функции Ф ищем ближайшее к 0,4. Это число 0,445, соответствующее значению аргумента 1,6. Поэтому

$$-C = 1,6.$$

Ответ: $C = -1,6$.

Пример 4. Вес пачек с рисом, расфасованных фирмой "Новый рис", является нормально распределенной случайной величиной с математическим ожиданием 1 кг и стандартным отклонением 10 г. Проверка

показала, что 2,5% выпускаемых пачек имеет вес меньше минимально допустимого стандартом. Каков этот минимальный вес?

Решение. Вес пачки риса является нормально распределенной величиной $X = \mathcal{N}(1000, 10)$. Для ответа на вопрос достаточно решить уравнение

$$p(X < C) = 0,025.$$

Преобразуем его

$$p\left(U < \frac{C - 1000}{10}\right) = 0,025.$$

Ясно, что

$$\frac{C - 1000}{10} < 0.$$

С учетом этого получаем

$$0,5 - \Phi\left(\frac{1000 - C}{10}\right) = 0,025,$$

$$\Phi\left(\frac{1000 - C}{10}\right) = 0,475,$$

$$\frac{1000 - C}{10} = 2,$$

$$C = 980.$$

Ответ: Минимально допустимый стандартный вес составляет 980 г.

7.2. Функция распределения случайной величины

Для читателей, хорошо усвоивших раздел, посвященный математическому анализу и понимающих смысл формулы Ньютона—Лейбница, можно дать рецепт, позволяющий просто и безошибочно считать вероятности вида

$$\int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

По формуле Ньютона—Лейбница подобные интегралы можно считать по любой первообразной для функции $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, в частности по $\Phi(x)$,

если только задать ее на всей числовой оси. Последнее сделать нетрудно, как было показано выше (см. пример 1 и последующие упражнения в третьей главе второй части): надо продолжить функцию нечетным образом, положив

$$\Phi(-x) = -\Phi(x).$$

Поскольку

$$\lim_{x\to-\infty} \Phi(x) = -\frac{1}{2} \text{ и } \lim_{x\to\infty} \Phi(x) = \frac{1}{2},$$

условимся считать $\Phi(-\infty) = -1/2$ и $\Phi(\infty) = 1/2$.

Тогда в конечных или бесконечных пределах

$$\int\limits_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \Phi(b) - \Phi(a).$$

Среди всех первообразных для произвольной случайной величины Y, имеющей плотность распределения $q(x)$, в теории вероятностей выделяют одну, которая называется функцией распределения. Она задается формулой

$$\int\limits_{-\infty}^{x} q(x)\, dx.$$

Другими словами, это та первообразная $F(x)$, для которой

$$\lim_{x\to-\infty} F(x) = 0$$

и, следовательно,

$$\lim_{x\to\infty} F(x) = 1$$

(рис. 7.11). Такая первообразная имеет простую интерпретацию:

$$F(x) = P(Y \leqslant x),$$

т.е. значение функции распределения в точке x равно вероятности попадания данной случайной величины в область левее x. Ясно, что чем правее

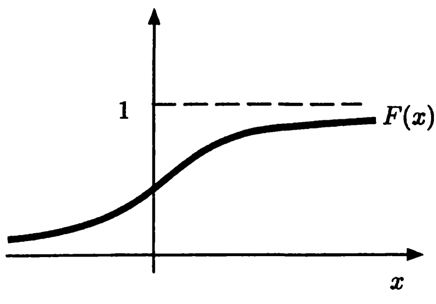


Рис. 7.11

точка x, тем шире область и тем больше вероятность для случайной величины попасть в эту область — следовательно, функция распределения, по крайней мере, не убывает. К этому же выводу приводит нас другое соображение: производная от $F(x)$ равна $q(x)$, а плотность распределения функция неотрицательная.

Условившись считать $F(-\infty) = 0$ и $F(\infty) = 1$, мы можем вычислять любые вероятности по формуле Ньютона—Лейбница

$$P(a < X \leqslant b) = F(b) - F(a)$$

(для непрерывных случайных величин вместо "$<$" можно писать "$\leqslant$" и наоборот).

Функцию распределения можно определить и для дискретных случайных величин по той же формуле, которая интерпретирует в терминах вероятности функцию распределения непрерывной случайной величины:

$$F(x) = P(Y \leqslant x),$$

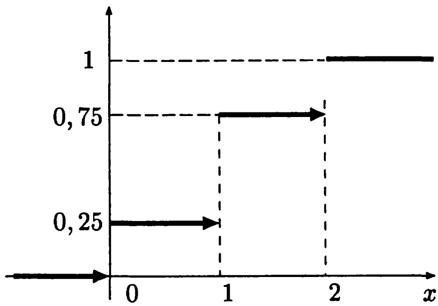


Рис. 7.12

Пример 5. Построим функцию распределения для биномиальной случайной величины $B = B(0{,}5,\ 2)$ (рис. 7.12).

Левее нуля у случайной величины значений нет, поэтому $F(x) = 0$. В точке 0 функция делает скачок, и

$$F(0) = P(B \leqslant 0) = P(B = 0) = 0{,}25.$$

Это же значение $F(x)$ сохраняет для всех точек интервала (0; 1), поскольку, например,

$$F(0{,}5) = P(B \leqslant 0{,}5) = P(B = 0) = 0{,}25.$$

Следующий скачок функция делает в точке $x = 1$, поскольку

$$P(B \leqslant 1) = P(B = 0) + P(B = 1) = 0{,}25 + 0{,}5 = 0{,}75.$$

В точке $x = 2$ функция делает последний скачок и $F(x) = 1$ при $x \geqslant 2$, поскольку все три возможных значения случайной величины 0, 1 и 2 оказываются левее каждой из точек $x \geqslant 2$.

Пользуясь функцией распределения, можно считать вероятности по формулам, аналогичным расмотренным выше. Например, для $B = B(0{,}5,\, 2)$

$$P(a < B \leqslant b) = F(b) - F(a),$$

но здесь знаки неравенств заменять уже нельзя.

Действительно, $F(2) - F(1) = 0{,}25 = P(B = 2)$, поэтому в выражении $P(1 < B \leqslant 2)$ нельзя менять знаки строгих неравенств на нестрогие и наоборот, чтобы не исключить нужное значение случайной величины (в данном случае 2) и. не включить лишние (в данном случае 1).

Польза функции распределения состоит прежде всего в том, что позволяет применять одни и те же методы как к непрерывным, так и к дискретным случайным величинам.

7.3. Формула Муавра—Лапласа

Напомним, что биномиальное распределение $B(n,p)$ имеет два параметра — n и p. Если случайная величина X распределена по закону $B(n,p)$, то это означает следующее:

$$p(X = k) = C_n^k p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

При этом

$$MX = np, \qquad DX = np(1-p).$$

Между биномиальным и нормальным распределениями существует простая связь. При больших значениях n справедлива следующая приближенная формула:

$$B(n,p) \approx N\left(np, \sqrt{np(1-p)}\right).$$

Эта формула называется *формулой Муавра—Лапласа*. Мы будем применять ее для вычисления вероятностей, связанных с биномиальным распределением, при выполнении условий

$$n > 90, \qquad np(1-p) > 9.$$

Пример 6. Игральный кубик бросают 120 раз. Какова вероятность того, что число выпадений "единицы" будет лежать в пределах от 20 до 30?

Решение. Здесь мы имеем дело с биномиальной случайной величиной $B\left(120, \frac{1}{6}\right)$. Проверим, можно ли применить формулу Муавра–Лапласа:

$$n = 120 > 90,$$
$$np(1-p) = 120 \cdot \frac{1}{6} \cdot \frac{5}{6} \approx 16{,}7 > 9.$$

Условия выполнены. Поэтому случайную величину $B(120, \frac{1}{6})$ можно заменить нормальным распределением $N(\mu, \sigma)$ с параметрами

$$\mu = np = 120 \cdot \frac{1}{6} = 20,$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{120 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 4,1.$$

Имеем:

$$p\left(20 \leqslant B\left(120, \frac{1}{6}\right) \leqslant 30\right) \approx p\,(20 \leqslant N(20; 4{,}1) \leqslant 30) =$$
$$= p\left(\frac{20-20}{4{,}1} \leqslant U \leqslant \frac{30-20}{4{,}1}\right) = p(0 \leqslant U \leqslant 2{,}4) = \Phi(2{,}4) = 0{,}492.$$

Пример 7. На некотором производстве регулярно проводится проверка качества продукции, для чего выбираются случайным образом 500 изделий и проверяются. Опыт показывает, что в среднем 11 изделий оказываются бракованными. Найти вероятность того, что во время очередной проверки число бракованных изделий окажется не менее 20.

Решение. Число бракованных изделий является биномиальной случайной величиной с параметрами $n = 500$ и $p = \frac{11}{500} = 0{,}022$. Проверим условия применимости формулы Муавра—Лапласа. Имеем

$$n = 500 > 90;$$
$$np(1-p) = 11 \cdot (1 - 0{,}022) = 10{,}758 > 9.$$

Условия выполнены, поэтому величину $B(500; 0{,}022)$ можно заменить нормальной случайной величиной с параметрами

$$\mu = np = 11,$$
$$\sigma = \sqrt{np(1-p)} = 3{,}3.$$

Имеем

$$p\,(B(500;0{,}022) \geqslant 20) \approx p\,(N(11;3{,}3) \geqslant 20) =$$

$$= p\left(U \geqslant \tfrac{20-11}{3{,}3}\right) = p(U \geqslant 2{,}7) = 0{,}5 - \Phi(2{,}7) =$$

$$= 0{,}5 - 0{,}497 = 0{,}003.$$

Замечание 2. (Центральная предельная теорема.) Формула Муавра—Лапласа является частным случаем следующего удивительного утверждения:
сумма большого числа почти произвольных случайных величин распределена приближенно по нормальному закону.

Этот факт является теоретическим обоснованием того, что многие реально наблюдаемые величины, испытывающие влияние множества случайных факторов, распределены по нормальному закону.

Упражнение 7.1. Монету бросают 5 раз. Случайная величина X — число выпадений герба. Составьте таблицу распределения, найдите математическое ожидание и дисперсию случайной величины X.

Ответ: 2,5; 1,25.

Упражнение 7.2. В лотерее на каждые 100 билетов приходится 15 выигрышей. Количество и размер выигрышей заданы в таблице.

Размер выигрыша	20	5	1
Количество выигрышей	1	4	10

Требуется составить закон распределения случайной величины — размера выигрыша в лотерее, приходящегося на один билет, а также найти математическое ожидание и дисперсию этой случайной величины.

Ответ: 0,5; 4,85.

Упражнение 7.3. Найдите вероятности (здесь и в следующей задаче $U = N(0,1)$ — стандартная нормальная случайная величина):

а) $p(0 \leqslant U \leqslant 0,7)$;
б) $p(U \geqslant 0,7)$;
в) $p(0,3 \leqslant U \leqslant 0,4)$;
г) $p(-0,1 \leqslant U \leqslant 0,2)$;
д) $p(-1,9 \leqslant U \leqslant -1,3)$;
е) $p(U \geqslant -2)$.

Ответ: а) 0,258; б) 0,242; в) 0,037; г) 0,119; д) 0,068; е) 0,977.

Упражнение 7.4. Найтдите число x такое, что

$$\text{а) } p(0 \leqslant U \leqslant x) = 0,4;$$
$$\text{б) } p(|U| < x) = 0,95.$$

Ответ: а) 1,3; б) 2.

Упражнение 7.5. Найдите следующие вероятности:

$$\text{а) } p(0 \leqslant N(1,3) \leqslant 2);$$
$$\text{б) } p(N(-20,10) < -10);$$
$$\text{в) } p(11,3 \leqslant N(15,7) \leqslant 16);$$
$$\text{г) } p(3 \leqslant N(0,30) \leqslant 9);$$
$$\text{д) } p(N(37,37) > 100).$$

Ответ: а) 0,236; б) 0,841; в) 0,232; г) 0,078; д) 0,045.

Упражнение 7.6. Производительность труда рабочих некоторого цеха является нормально распределенной случайной величиной с математическим ожиданием 90 кг за смену и стандартным отклонением 15 кг за смену. Вычислите долю рабочих, производительность которых:

а) лежит в промежутке от 80 до 110 кг за смену;

б) превышает 110 кг за смену;

в) менее 80 кг за смену.

г) Какой следует установить норму дневной выработки, чтобы 90% рабочих ее выполняли?

Ответ: а) 0,66; б) 0,1; в) 0,24; г) 70,5 кг.

Упражнение 7.7. Производство дает 1% брака. Какова вероятность того, что из взятых на исследование 1100 изделий забраковано будет не более 17?

Ответ: 0,964.

Упражнение 7.8. Какова вероятность того, что при 100-кратном бросании монеты число выпадений герба будет от 45 до 55?

Ответ: 0,682.

Глава 8

Случайные величины (окончание)

8.1. Математическое ожидание и дисперсия биномиальной случайной величины

Как уже отмечалось, основное преимущество дисперсии как меры разброса случайной величины перед другими мерами состоит в том, что дисперсия суммы независимых случайных величин равна сумме дисперсий случайных величин-слагаемых. Пользуясь теоремой о сложении дисперсий, легко вычислить дисперсию биномиальной случайной величины $B(n,p)$.

Рассмотрим случайную величину $B(1,p)$, которая, как мы знаем, задается таблицей

0	1
q	p

Ее математическое ожидание $MB(1,p)$ равно $0 \cdot q + 1 \cdot p = p$.

Вычисляем теперь дисперсию. Поскольку $1 - p = q$, то

$$DB(p,1) = (0-p)^2 q + (1-p)^2 p = p^2 q + q^2 p = pq(p+q) = pq.$$

Теперь по теоремам о сложении математического ожидания и дисперсии мы легко узнаем эти числовые параметры любой биномиальной случайной величины. Для этого достаточно понять, что $B(n,p)$ является n-кратной суммой независимых испытаний $B(1,p)$, и, следовательно, ее математическое ожидание равно $n \cdot p$, а дисперсия равна $n \cdot pq$.

8.2. Неравенство Чебышева

Если дисперсия случайной величины, имеющей нулевое математическое ожидание, равна 1, то значение 10 эта случайная величина не может принимать с вероятностью, большей, чем 0,01. Действительно, если бы значению 10 соответствовала, например, вероятность $p_{10} = 0{,}1$, то в сумме, составляющей дисперсию этой случайной величины, имелось бы слагаемое $10^2 \cdot 0{,}1$, которое равно 10 — а это значит, что ее дисперсия должна была оказаться еще больше (поскольку сумма содержит еще какие-то неотрицательные слагаемые).

Эта простая идея позволяет доказать элементарную теорему, следствия из которой имеют огромную важность.

Теорема 8.1. ***(Неравенство Чебышева.)*** *Для любой случайной величины X вероятность того, что она отклонится от своего математического ожидания больше, чем на число α, всегда меньше, чем DX/α^2. В виде формулы утверждение записывается так*[1]:

$$P(|X - MX| \geqslant \alpha) \leqslant \frac{DX}{\alpha^2}.$$

Доказательство. Проведем доказательство в более наглядном случае дискретной случайной величины. В случае непрерывной величины можно провести аналогичные рассуждения с помощью интегралов или аккуратно воспользоваться дискретизацией, которую мы ввели в главе 6.

Итак, пусть случайная величина X принимает значения из набора $x_1, x_2, \ldots, x_n$ с соответствующими вероятностями $p_1, p_2, \ldots, p_n$, причем эти значения пронумерованы по возрастанию удаленности от математического ожидания MX (если это не так, их всегда можно перенумеровать).

Отметим номер k, начиная с которого расстояние $|x_k - MX|$ становится больше или равным α. Тогда

$$DX = (x_1 - MX)^2 p_1 + \ldots + (x_n - MX)^2 p_n \geqslant (x_k - MX)^2 p_k + \ldots + (x_n - MX)^2 p_n,$$

поскольку в последней сумме мы рассматриваем лишь часть слагаемых. Далее заметим, что в этой сумме содержатся лишь слагаемые, для

[1] В словесной формулировке мы употребили "больше" и "меньше" вместо "больше или равно" и "меньше или равно" для более легкого чтения. Утверждение верно в обеих формах, но обычно употребляется во второй.

которых $|x_k - MX| \geqslant \alpha$. Заменив все квадраты выражений в скобках на меньшее либо равное число α^2, мы еще уменьшим сумму либо оставим ее без изменения. Таким образом,

$$DX \geqslant \alpha^2 p_k + \ldots + \alpha^2 p_n = \alpha^2 (p_k + \ldots + p_n).$$

Осталось заметить только, что $p_k + \ldots + p_n$ есть сумма вероятностей тех значений, которые отклоняются от MX больше, чем на α, т.е. вероятность того, что случайная величина будет иметь это большое отклонение. Окончательно

$$DX \geqslant \alpha^2 \cdot P(|X - MX| \geqslant \alpha).$$

Деля обе части на α^2, получаем утверждение теоремы.

Как мы установили, дисперсия биномиальной случайной величины B(p,n) равна npq. Предположим, мы бросаем симметричную монетку 100 раз. Тогда вероятность того, что суммарное количество "гербов" будет меньше 40 или больше 60, не превосходит $100pq/10^2 = 100 \cdot 0{,}5 \cdot 0{,}5/100 = 1/4$.

Если мы бросим монетку 10 000 раз и рассмотрим вероятность пропорционального отклонения на 1/10 возможного диапазона, т.е. на 1000, то вероятность окажется совершенно ничтожной: $10000 \cdot 0{,}5 \cdot 0{,}5/1000^2 = 1/400$, а 1/4 будет равна вероятность отклонения на 100, т.е. на $\sqrt{10\,000}$.

Это значит, что с вероятностью 3/4 количество "гербов" отклонится от 1/2 меньше, чем на 1%.

Если же бросить монету 1 000 000 раз, то с вероятностью 3/4 количество "гербов" отклонится от 1/2 меньше, чем на 0,1%.

Мы, таким образом, обосновали нашу веру в то, что увеличение количества испытаний приводит к приближению наблюдаемой частоты к теоретической вероятности.

Неравенство Чебышева позволяет также понять, почему в статистике чаще используется не дисперсия, а среднее квадратическое отклонение — корень из дисперсии. Пусть случайная величина X имеет дисперсию DX. Положим в неравенстве Чебышева $\alpha = 3\sqrt{DX}$. Тогда

$$P(|X - MX| \geqslant 3\sqrt{D}) \leqslant \frac{DX}{9DX} = \frac{1}{9}.$$

Это значит, что для любой случайной величины отклонение от среднего значения на три среднеквадратических отклонения происходит с

вероятностью, не превосходящей 1/9. Отклонение больше, чем на четыре среднеквадратических отклонения, как легко подсчитать, имеют вероятность не больше 1/16 и т.д. Среднее квадратическое отклонение обозначается обычно символом σ_X. То, что постоянные вероятности соответствуют отклонениям, "измеренным" именно в масштабе σ_X, делает этот масштаб весьма удобным для оценки "расстояния" между результатами испытаний случайных величин. Если $\sigma_X = 1$, то разность в 1,5 единицы между двумя испытаниями случайной величины X будет в каком-то смысле (который разъяснится позже) столь же вероятной, как и разность в 150 единиц между двумя испытаниями случайной величины Y, имеющей среднее квадратическое отклонение $\sigma_Y = 100$.

8.3. Закон больших чисел

Последнее употребление, которое мы дадим неравенству Чебышева — это доказательство того факта, что среднее арифметическое результатов независимых испытаний случайной величины при увеличении числа испытаний все с большей точностью представляет математическое ожидание данной случайной величины.

Пусть X случайная величина, имеющая математическое ожидание M и дисперсию D. Представим ситуацию следующим образом. Пусть $X_1, \ldots, X_n$ случайные величины, подобные X, т.е. имеющие все те же самые параметры распределения. (Например, в случае бросания монет, мы можем заготовить n одинаковых симметричных монет.) Возьмем теперь новую случайную величину

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

Ее однократное испытание — это то же самое, что последовательное испытание $X_1, \ldots, X_n$ и затем вычисление среднего арифметического полученных результатов. Математическое ожидание $\bar{X}$ есть сумма математических ожиданий $MX_i = M$, деленная на n, т.е. $Mn/n = M$.

Что касается дисперсии, то тут ситуация такова: по теореме сложения дисперсий

$$D(X_1 + \ldots + X_n) = nD,$$

но по теореме о дисперсии случайной величины, умноженной на константу,

$$D(\bar{X}) = D\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{D(X_1 + \ldots + X_n)}{n^2} = \frac{D}{n}.$$

Теперь мы можем применить неравенство Чебышева к случайной величине $\bar{X}$.

$$P(|\bar{X} - M| \geqslant \alpha) \leqslant \frac{D}{n\alpha^2}.$$

Отсюда следует, что

$$P(|\bar{X} - M| \leqslant \alpha) \geqslant 1 - \frac{D}{n\alpha^2}.$$

Следствием последнего неравенства, которое называется *законом больших чисел в форме Чебышева*, является утверждение: для любого как угодно малого α

$$\lim_{n\to\infty} P(|\bar{X} - M| \leqslant \alpha) = 1.$$

Среднее арифметическое результатов испытаний с ростом n все точнее отражает математическое ожидание испытываемой случайной величины.

Часть IV

Математическая статистика

Глава 1

Первичная обработка и точечные оценки

По словам И. Канта, в естественной науке ровно столько собственно науки, сколько в ней математики. Психология хотя и не принадлежит к естественным наукам, но и не вовсе чужда им. Ее специфическое положение устанавливает особые отношения эмпирических психологических исследований с математическими методами обработки данных. Поскольку данные являются массивами чисел, говорить о них "больше", "меньше" и т.д. можно только в рамках хорошо проработанной науки — математической статистики. Поскольку числа эти отражают тонкие и неповторимые моменты психических явлений, выводы о числах, полученные на основании математической обработки, не приложимы непосредственно к описываемой ими специфической реальности, а требуют аккуратной и даже в определенной степени критически к себе настроенной интерпретации.

В идеале эта интерпретация должна опираться не только на слой выводов, поставляемых компьютерными статистическими программами, но и на понимание сути проделываемых в "черном ящике" статистического пакета операций и преобразований. Чем богаче у психолога представление о математической сути применяемых методов, тем яснее его понимание собственных результатов.

Авторы полагают, что предыдущие главы книги в достаточной степени подготовили читателя к пониманию этой сути.

Как и прежде, нечетные главы можно читать, не заглядывая в четные. Для требовательного читателя в четных главах рассматриваются

некоторые более сложные методы работы с данными и даются объяснения сути изложенного в нечетных главах материала.

1.1. Первичная обработка данных

Обычно полученные в результате наблюдений результаты представляют собой набор чисел. Просматривая этот набор, как правило, трудно выявить какую-либо закономерность. Поэтому данные подвергают некоторой первичной обработке, целью которой является упрощение дальнейшего анализа. Мы рассмотрим подробно один из возможных способов.

Рассмотрим данные, полученные в результате регистрации значений некоторой случайной величины — набор чисел

$$x_1, x_2, \ldots, x_n$$

(отметим, что некоторые значения могут совпадать). Этот набор чисел называется *выборкой.*

Дальнейшие действия зависят от того, насколько много в выборке *различных* чисел. Если мы имеем дело с дискретной случайной величиной, то различных чисел немного; если с непрерывной случайной величиной, то, скорее всего, все числа окажутся различными. Поэтому далее рассмотрим два этих случая по отдельности.

Дискретный случай

Первый этап обработки выборки — это составление *вариационного ряда*. Его получают так — среди всех чисел x_i отбирают все различные и располагают в порядке возрастания:

$$\alpha_1, \alpha_2, \ldots, \alpha_m,$$

где $\alpha_1 < \alpha_2 < \ldots < \alpha_m$.

Следующий этап обработки выборки — составление *дискретной таблицы частот*:

α_1	α_2	...	α_m
k_1	k_2	...	k_m
$n_1 = k_1/n$	$n_2 = k_2/n$	...	$n_m = k_m/n$

Здесь n — число всех измерений, k_i — число измерений, в которых наблюдалось значение a_i. Величины k_i называются *частотами*, а величины $n_i = k_i/n$ — *относительными частотами*.

Графической иллюстрацией дискретной таблицы частот является *столбиковая диаграмма* (рис. 1.1).

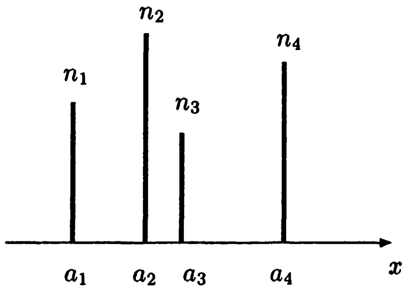


Рис. 1.1

Замечание 1. В дискретной таблице частот частоты и относительные частоты пропорциональны. Поэтому при построении столбиковой диаграммы на вертикальной оси можно указывать значения либо относительных частот, либо частот — визуальное восприятие от этого не зависит.

Пример 1. Пусть нашей задачей является выявление картины успеваемости студентов, сдавших экзамен по курсу "Специальная психология". Курс прослушало 56 человек. Полученные студентами оценки представляют собой (в порядке алфавитного списка) следующий набор чисел
3, 4, 5, 4, 3, 3, 5, 4, 3, 5, 5, 2, 3, 5, 3, 5, 3, 5, 4, 4, 3, 3, 4, 3, 4, 3, 3, 5, 3, 3, 4, 3, 4, 3, 5, 3, 4, 4, 3, 5, 3, 3, 5, 4, 2, 5, 3, 4, 2, 3, 5, 4, 3, 5, 3, 5.

Это и есть исходные данные — выборка. Числа, составляющие выборку, представляют собой реализации случайной величины — оценки на экзамене.

Составление вариационного ряда не представляет сложностей. Вот он:

$$2, 3, 4, 5.$$

Теперь надо подсчитать, сколько раз встречается каждая из оценок. Таблица частот выглядит следующим образом:

2	3	4	5
3	24	14	15
$\frac{3}{56} \approx 0{,}05$	$\frac{24}{56} \approx 0{,}43$	$\frac{14}{56} \approx 0{,}25$	$\frac{15}{56} \approx 0{,}27$

Здесь в последней строке — относительные частоты, получающиеся при делении частот на число измерений $n = 56$.

Столбиковая диаграмма, иллюстрирующая полученную таблицу, изображена на рис. 1.2.

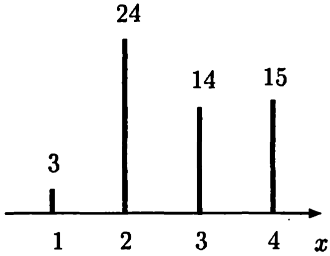


Рис. 1.2

Непрерывный случай

Если число различных значений в выборке велико, вычислять частоту каждого из них не имеет большого смысла. Например, если *все* значения в выборке различны, то при попытке составить дискретную таблицу частот получается вот что:

α_1	α_2	...	α_m
1	1	...	1
$1/n$	$1/n$	...	$1/n$

Понятно, что такая таблица не добавляет наглядности.

Поэтому поступают следующим образом. Весь промежуток изменения значений выборки, от минимального до максимального, разбивают на интервалы. После этого подсчитывают число значений из выборки, попадающих в каждый интервал (частоты), а затем — относительные частоты. В результате получаем *интервальную таблицу частот*:

$[\mu_1; \mu_2]$	$(\mu_2; \mu_3]$	...	$(\mu_m; \mu_{m+1}]$
k_1	k_2	...	k_m
$n_1 = k_1/n$	$n_2 = k_2/n$	...	$n_m = k_m/n$

Здесь n — число всех измерений, m — число интервалов, k_i — количество чисел, приходящихся на i-й интервал, $n_i = k_i/n$ — относительная частота попадания в i-й интервал. Интервалы обычно берут одинаковой длины, хотя это и не обязательно.

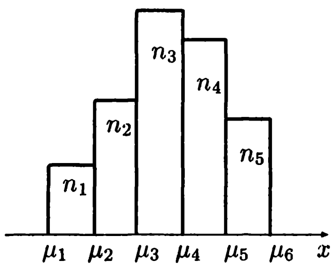


Рис. 1.3

Графической иллюстрацией интервальной таблицы частот является *гистограмма* (рис. 1.3). Гистограмма представляет собой ступенчатую линию; основанием i-й ступеньки является интервал $(\mu_i; \mu_{i+1}]$, а площадь этой ступеньки равна n_i.

Замечание 2. Число интервалов m выбирают из соображений наглядности получающейся гистограммы. Обычно m лежит в пределах от 5 до 15.

Замечание 3. Если интервалы $(\mu_i; \mu_{i+1}]$ выбраны одинаковой длины, то площади ступенек гистограммы пропорциональны их высотам. В этом случае можно отмечать на оси ординат просто частоты k_i.

Пример 2. Предположим, что студенты некоторой группы, состоящей из 25 человек, написали контрольную работу. Каждый студент получил определенное количество баллов. Приведем эти баллы (в порядке алфавитного списка группы):

75, 145, 150, 180, 125, 150, 150, 165, 95, 135, 130, 70, 130, 105, 135, 135, 100, 160, 60, 85, 120, 60, 145, 150, 135.

Требуется построить интервальную таблицу частот и гистограмму.

Нетрудно найти среди приведенных чисел минимальное и максимальное — это числа 60 и 180. Таким образом, все значения лежат в отрезке [60; 180]. Разобьем этот отрезок, например, на $m = 6$ равных частей. После этого подсчитаем число значений, попавших в каждый интервал (воспользуемся методом, описанным в примере 1):

[60; 80]: 4 значения
(80; 100]: 3 значения
(100; 120]: 2 значения
(120; 140]: 7 значений
(140; 160]: 7 значений
(160; 180]: 2 значения

Построим теперь интервальную таблицу частот:

[60; 80]	(80; 100]	(100; 120]	(120; 140]	(140; 160]	(160; 180]
4	3	2	7	7	2
$\frac{4}{25} = 0{,}16$	$\frac{3}{25} = 0{,}12$	$\frac{2}{25} = 0{,}08$	$\frac{7}{25} = 0{,}28$	$\frac{7}{25} = 0{,}28$	$\frac{2}{25} = 0{,}08$

Соответствующая гистограмма изображена на рис. 1.4. На вертикальной оси проставлены частоты — см. замечание 3.

При взгляде на полученную гистограмму можно сделать вывод, что бóльшая часть группы подготовилась к контрольной на довольно высоком уровне (интервал (120; 180]). Другая часть, поменьше, подготовилась плохо (интервал [60; 100]) и совсем мала группа с промежуточным баллом ([100; 120]).

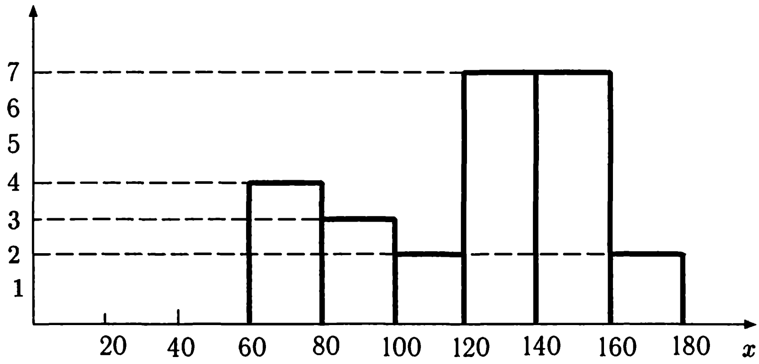


Рис. 1.4

1.2. Точечные оценки

Выше были рассмотрены некоторые методы первичной обработки данных, в ходе которой имеющиеся "сырые" результаты наблюдений преобразовываются для достижения большей наглядности. Теперь мы рассмотрим методы, позволяющие, исходя из тех же данных, делать предположения относительно числовых характеристик наблюдаемой случайной величины — математическом ожидании и дисперсии.

В связи с этим возникает задача: исходя из набора значений (выборки)

$$x_1, x_2, \ldots, x_n$$

величины X, полученного в результате n независимых наблюдений, оценить значение математического ожидания MX, дисперсии DX либо еще какого-нибудь параметра.

Повторим сказанное несколько иными словами. Имеется случайная величина X, значения (реализации) которой $x_1, x_2, \ldots, x_n$ каким-либо образом становятся нам известными. (Этой случайной величиной может быть число посетителей данного магазина в течение дня, рост в сантиметрах студента данного вуза, годовой доход гражданина данной страны и пр.) У величины X имеется, скажем, математическое ожидание, которое нам неизвестно. Требуется найти способ, при помощи которого по известным реализациям величины X можно разумным образом оценить неизвестное математическое ожидание.

Пример 3. Пусть случайная величина X это рост первого пассажира метро, входящего на некую станцию в произвольное воскресенье после 11.30.

Мы уже говорили об операциях над случайными величинами, в частности, о сумме случайных величин и об умножении случайной величины на константу. Рассмотрим следующее выражение, составленное из одинаковых случайных величин, имеющих одинаковое распределение, совпадающее с нашей X:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{15}}{15}.$$

Это выражение $\bar{X}$ представляет собой случайную величину, принимающую определенное значение, если мы в течение 15 воскресных дней в 11.30 измерим рост первого входящего пассажира, а затем вычислим среднее арифметическое полученных результатов. Ее закон распределения будет зависеть от закона распределения случайной величины X (последнему подчинена каждая из величин $X_1, X_2, \ldots, X_{15}$).

Точно так же для произвольного n

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Случайные величины $X_1, X_2, \ldots, X_n$ имеют один и тот же закон распределения, совпадающий с законом распределения величины X. Поэтому

$$M\bar{X} = M\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n}(MX_1 + \ldots + MX_n) = \frac{1}{n}nMX = MX.$$

$M\bar{X}$ при внимательном рассмотрении оказывается средним ожидаемым результатом комплексного испытания "измерение среднего роста n воскресных пассажиров".

В таком случае последнее равенство отражает интуитивно очевидный факт: в среднем наша оценка ожидаемого роста воскресного пассажира с помощью среднего роста n испытуемых будет колебаться вокруг реального среднего роста пассажиров.

Можно показать, что эти колебания будут все более узкими по мере увеличения n. Действительно, рассмотрим дисперсию случайной величины $\bar{X}$ = {средний рост n пассажиров}.

$$D\bar{X} = D\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2}D(X_1 + \ldots + X_n) = \frac{1}{n^2}nDX = \frac{DX}{n}.$$

Чем больше n, тем меньше дисперсия, характеризующая разброс нашей оценки среднего роста вокруг реального среднего роста. Это

значит, что среднее арифметическое роста n испытуемых

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

можно рассматривать как оценку среднего роста людей, которых эти испытуемые представляют. Оценку тем более точную, чем больше n.

Вопрос об оценке разброса роста пассажира, а именно об оценке дисперсии случайной величины X, оказывается не столь простым

При тех же обстоятельствах для оценки дисперсии случайной величины X = {рост случайного пассажира} служит формула

$$S_x^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2).$$

Замечание 4. Дисперсия случайной величины X обозначается DX. Сочетанием S_x^2 обозначают обычно оценку дисперсии, произведенную по выборке $(x : x_1, \ldots, x_n)$ по данной выше формуле. Оценка среднеквадратического отклонения задается формулой $\sqrt{S_x^2}$ и обозначается s_x.

Дисперсию можно оценивать и некоторыми другими формулами. За каждым таким способом оценивания закрепляются обычно какие-то стандартные обозначения.

Почему для оценки дисперсии берется сумма квадратов отклонений роста испытуемых от среднего роста, мы объясним в следующей главе. Обратим внимание на множитель $\frac{1}{n-1}$. Сумма в скобках содержит n слагаемых, и, казалось бы, естественно было делить сумму так же, как и в случае математического ожидания, на n, т.е. вычислить среднее арифметическое квадратов отклонений. Не вдаваясь в подробные объяснения, укажем, что в среднем такая оценка дала бы заниженный показатель. Деля на $n - 1$ мы, как ни странно, получаем в среднем точную оценку дисперсии.

В следующих разделах мы разберем иные оценки математического ожидания, которые в психологических экспериментах обладают некоторыми преимуществами, по сравнению со средним арифметическим. Для любых таких оценок важны две характеристики: *состоятельность* и *несмещенность*. Первая означает, что при увеличении размера выборки оценка становится все более точной, вторая -- что по выборке любого размера оценка дает в среднем правильный результат (ее математическое ожидание равно оцениваемому параметру случайной величины).

Упражнение 1.1. Как нам уже известно, оценка дисперсии

$$S_x^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)$$

состоятельная и несмещенная. Докажите, что оценка

$$\frac{1}{n}((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)$$

не является несмещенной (вычислите ее математическое ожидание, исходя из того, что $MS_X^2 = DX$). Пользуясь своими практическими навыками в вычислении пределов, докажите, что она состоятельна.

1.3. Оценки вероятности события

Пусть некоторое событие происходит в результате единичного испытания с вероятностью P, которая нам неизвестна. Однако ситуация позволяет многократно повторить испытание и подсчитать, сколько раз произошло указанное событие. Более точно: пусть произведено n испытаний, в которых событие произошло k раз. Здесь, в отличие от предыдущих рассмотрений, исходными данными для анализа будут всего два числа — n и k.

Задача, которую мы будем рассматривать, заключается в отыскании оценки неизвестной вероятности P по имеющимся данным n, k.

Из соображений здравого смысла ясно, что точечная оценка $\tilde{P}$ вероятности P определяется следующим соотношением

$$\tilde{P} = \frac{k}{n}. \tag{1}$$

Докажем несмещенность и состоятельность этой оценки.

При любом фиксированном n величина k является случайной величиной с биномиальным законом распределения

$$k = B(n, P).$$

Напомним, что в этом случае

$$Mk = nP, \qquad Dk = nP(1 - P).$$

Докажем несмещенность оценки $\tilde{P}$. Имеем

$$M\tilde{P} = M\left(\frac{k}{n}\right) = \frac{1}{n}Mk = \frac{1}{n}nP = P.$$

Теперь вычислим величину $D\tilde{P}$:

$$D\tilde{P} = D\left(\frac{k}{n}\right) = \frac{1}{n^2}Dk = \frac{1}{n^2}nP(1-P) = \frac{P(1-P)}{n}.$$

Таким образом,

$$D\tilde{P} = \frac{P(1-P)}{n} \longrightarrow 0 \qquad \text{при} \qquad n \longrightarrow \infty,$$

что доказывает состоятельность точечной оценки $\tilde{P}$.

Глава 2

Плотности, гистограммы и выборочные оценки параметров распределения

2.1. Почему непохожие формулы выражают одно и то же

В этой главе будет разъяснена связь между уже введенными понятиями. В главе 6 предыдущей, третьей части книги мы уже проясняли связь формул математического ожидания и дисперсии в дискретном и непрерывном случаях. Продолжая эту тему, мы рассмотрим здесь формулы математического ожидания и дисперсии, формулы их выборочной оценки и гистограммы распределения, построенные по выборке.

В предыдущей главе мы доказали, что среднее арифметическое выборочных значений случайной величины оценивает ее математическое ожидание. Какова связь между этим способом оценивания и формулой математического ожидания непрерывной случайной величины?

Пусть случайная величина X имеет плотность распределения $p(x)$. Предположим также для удобства, что случайная величина принимает значения из отрезка от a до b, и даже конкретнее, от 0 до 10 (общий случай будет отличаться совершенно несущественно).

Разделим отрезок [0; 10] на 10 частей соответствующими целым числам точками. Рассмотрим первый отрезок между точками 0 и 1. Вероятность того, что значение случайной величины попадет в этот

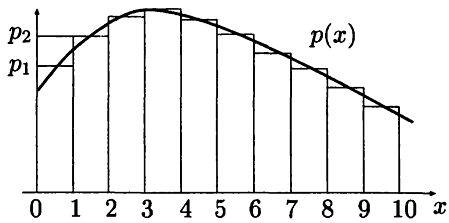


Рис. 2.1

отрезок, равную соответствующему интегралу от плотности $p(x)$, обозначим p_1. Если мы провели достаточно много независимых испытаний случайной величины X, то выборка результатов будет содержать приблизительно соответствующую p_1 долю значений, попавших в этот первый отрезок (мы доказали это в предыдущей главе). Предположим, что эта доля просто равна вероятности p_1, т.е. если выборка содержит n чисел, то в первом отрезке из них ровно np_1. Точно так же $np_2, np_3, \ldots, np_{10}$ — количество элементов выборки во втором, третьем и т.д. отрезках. Если мы построим гистограмму нашей выборки (рис. 2.1), то последнее предположение на рисунке отразится в равенстве площадей криволинейной трапеции под $p(x)$ с основанием [0; 1] и соответствующего прямоугольника гистограммы. Это значит, что на маленьком отрезке [0; 1] значение $p(x)$ мало отличается от высоты первого прямоугольника, равной p_1.

Упражнение 2.1. При разбиении оси абсцисс на единичные отрезки, высота прямоугольников гистограммы равна отношению k/n — количества элементов выборки, попавших в данный отрезок, к общему объему выборки. Если разбить ось абсцисс на отрезки длиной 0,5, что надо брать в качестве высоты прямоугольников, чтобы гистограмма по-прежнему была похожа на график плотности $p(x)$?

Заметим далее сходство рис. 2.1 с рис. 6.1 предыдущей части, с помощью которого устанавливалась связь между формулами математического ожидания в непрерывном и дискретном случаях. Интеграл в формуле математического ожидания можно разложить в сумму следующих десяти слагаемых

$$\int_0^{10} x\,p(x)\,dx = \int_0^1 x\,p(x)\,dx + \int_1^2 x\,p(x)\,dx + \ldots + \int_9^{10} x\,p(x)\,dx.$$

Рассмотрим, например, последний из них

$$\int_9^{10} x\,p(x)\,dx.$$

Подынтегральная функция $xp(x)$, взятая на маленьком отрезке, мало отличается от произведения среднего значения x на этом отрезке (равного 0,95) и среднего значения $p(x)$ на этом отрезке (равного p_{10}). Это значит, что[1]

$$\int_9^{10} x\,p(x)\,dx \approx 9{,}5 \cdot p_{10}.$$

Аналогично для других отрезков, поэтому

$$\int_0^{10} x\,p(x)\,dx \approx 0{,}5 \cdot p_1 + 1{,}5 \cdot p_2 \ldots + 9{,}5 \cdot p_{10}.$$

Рассмотрим теперь среднее арифметическое выборочных значений. Сгруппируем слагаемые в 10 групп, отнеся к каждой только те члены выборки, которые попадают в соответствующий отрезок.

$$\begin{array}{lcccl} \bar{X} = & & & & \\ (x_1 & + & \ldots & + \; x_{k_1} & + \\ +x_{k_1+1} & + & \ldots & + \; x_{k_2} & + \\ & & \ldots & & \\ +x_{k_9+1} & + & \ldots & + \; x_n &)/n. \end{array}$$

В первой строке содержатся k_1 элементов выборки, которые попали в первый отрезок. Их величина мало отличается от среднего значения — середины "их" отрезка, точки 0,5. Их количество k_1, деленное на n, равно по нашему предположению p_1, поэтому

$$(x_1 + x_2 + \ldots + x_{k_1})/n \approx 0{,}5 \cdot k_1/n = 0{,}5 \cdot p_1.$$

Аналогично для других строк. В целом это означает приближенное равенство

$$\bar{X} \approx 0{,}5 \cdot p_1 + 1{,}5 \cdot p_2 \ldots + 9{,}5 \cdot p_{10}.$$

Приблизительные, "гистограммные" представления формулы математического ожидания в виде интеграла и формулы выборочной оценки

[1] Площадь криволинейной трапеции под функцией $f(x)$ с основанием единичной длины, как это имеет место в нашем случае, равно просто некоторому среднему значению функции $f(x)$ — высоте равновеликого прямоугольника на том же единичном основании.

математического ожидания (как среднего арифметического) совпадают. Расхождения между всеми тремя видами формул тем меньше, чем больше размер выборки и чем мельче разбиение оси абсцисс.

Несколько сложнее истолковывается формула выборочной дисперсии. рассмотрим сначала более естественный вариант формулы

$$S_x^2 = \frac{1}{n}((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2),$$

отличающийся множителем $1/n$, что, как мы объясняли в предыдущей главе, дает несколько заниженную оценку дисперсии. Тем не менее при увеличении n эта ошибка стремится к нулю, поэтому приведенные ниже приблизительные равенства имеют смысл. Более подробно этой темы мы коснемся в следующем параграфе.

Обозначим математическое ожидание случайной величины X буквой a. Интеграл в формуле дисперсии разложим в сумму десяти слагаемых, как мы это делали для математического ожидания.

$$DX = \int_0^{10} (x-a)^2\, p(x)\, dx =$$

$$= \int_0^1 (x-a)^2\, p(x)\, dx + \int_1^2 (x-a)^2\, p(x)\, dx + \ldots + \int_9^{10} (x-a)^2\, p(x)\, dx.$$

Рассмотрим одно из них, например, последнее

$$\int_9^{10} (x-a)^2\, p(x)\, dx.$$

Здесь мы можем почти дословно повторить предыдущие рассуждения. Подынтегральная функция $(x-a)^2 p(x)$, взятая на маленьком отрезке, мало отличается от произведения приблизительно среднего значения $(x-a)^2$ на этом отрезке (равного $(9{,}5-a)^2$) и среднего значения $p(x)$ на этом отрезке (равного p_{10}). Это значит, что

$$\int_9^{10} (x-a)^2\, p(x)\, dx \approx (9{,}5-a)^2 \cdot p_{10},$$

т.е. примерно равен квадрату отклонения от математического ожидания середины отрезка, умноженному на вероятность попадания случайной величины в данный отрезок.

Аналогично для других отрезков, поэтому

$$DX = \int_0^{10} (x-a)^2\, p(x)\, dx \approx (0{,}5-a)^2 \cdot p_1 + (1{,}5-a)^2 \cdot p_2 \ldots + (9{,}5-a)^2 \cdot p_{10}.$$

Рассмотрим теперь нашу формулу выборочной оценки дисперсии. Сгруппируем слагаемые в 10 групп, отнеся к каждой только те члены выборки, которые попадают в соответствующий отрезок.

$$\begin{array}{rcccccl} S_x^2 = & & & & & & \\ ((x_1 - a)^2 & + & \ldots & + & (x_{k_1} - a)^2 & + & \\ +(x_{k_1+1} - a)^2 & + & \ldots & + & (x_{k_2} - a)^2 & + & \\ & & \ldots & & & & \\ +(x_{k_9+1} - a)^2 & + & \ldots & + & (x_n - a)^2 & &)/n. \end{array}$$

Сказанное выше о среднем арифметическом верно и в данном случае. В первой строке содержатся k_1 элементов выборки, которые попали в первый отрезок. Поскольку эти x_i мало отличается от среднего значения 0,5 — середины “их” отрезка, то $(x_i - a)^2$ приближенно равно $(0{,}5 - a)^2$. Как и прежде, k_1/n, равно p_1, поэтому

$$((x_1-a)^2+(x_2-a)^2+\ldots+(x_{k_1}-a)^2)/n \approx (0{,}5-a)^2 \cdot k_1/n = (0{,}5-a)^2 \cdot p_1.$$

Аналогично для других строк. В целом это означает приближенное равенство

$$S_x^2 \approx (0{,}5 - a)^2 \cdot p_1 + (1{,}5 - a)^2 \cdot p_2 \ldots + (9{,}5 - a)^2 \cdot p_{10}.$$

Снова приблизительные, “гистограммные” представления формул совпали.

2.2. О степенях свободы

Этот короткий раздел можно считать факультативным. В нем объясняется общее для многих разделов математики понятие “степени свободы”. Мы рекомендуем его чтение, во-первых, тем, кто хочет лучше понимать статистику, а во-вторых, тем немногочисленным читателям, в сферу интересов которых попадают вопросы освоения двигательных навыков.

Понятие “число степеней свободы” сформировалось в механике. Возьмем, например, вырезанный из картона треугольник. Понятно, что

его положение в пространстве целиком определяется положением трех его вершин, т.е. тремя тройками координат (x_1, y_1, z_1), (x_2, y_2, z_2) и (x_3, y_3, z_3).

Но можно заметить, что если первая вершина помещена в некоторой точке, то вторая не может удалиться от нее на расстояние, большее, чем длина соединяющего их ребра. Это значит, что задав две координаты второй вершины, мы можем вычислить положение третьей координаты. Третья же вершина, после того как заданы координаты двух предыдущих, вообще задается одной координатой, а две другие могут быть вычислены.

Таким образом, положение в пространстве картонного треугольника определяется шестью координатами, что и выражается словами "имеется шесть степеней свободы для задания положения треугольника". Это, однако, не самое важное.

Между вершинами треугольника имеются три соотношения:
— расстояние от первой до второй равно длине соответствующего ребра,
— расстояние от второй до третьей равно длине соответствующего ребра,
— расстояние от первой до третьей равно длине соответствующего ребра.

Число степеней свободы системы равно числу степеней свободы ее элементов минус число наложенных связей. В нашем случае число степеней свободы треугольника вычисляется $6 = 9 - 3$, где 9 — три тройки координат вершин, а 3 — число наложенных связей.

Наиболее близкий пример из психологии принадлежит Н.А. Бернштейну: если вам надо прикоснуться кончиком пальца к дверному звонку, то положение руки, обеспечивающей этот акт, имеет по крайней мере 7 степеней свободы (считаем положение плечевого сустава фиксированным, каждый из остальных 5 суставов добавляет две степени свободы минус три степени, заданные соотношением "кончик указательного пальца находится на звонке").

Мы упоминали о степенях свободы в первой части книги. Здесь мы можем пояснить сказанное с этих общих позиций. Точка в пространстве имеет три свободных координаты. Точка, подчиненная соотношению $ax + by + cz = 0$ (линейному уравнению) теряет одну степень свободы и может быть задана двумя параметрами, а две степени свободы имеет точка на плоскости.

Точка, координаты которой удовлетворяют двум соотношениям (системе двух уравнений), теряет две степени свободы и описывается од-

ним параметром, а одна степень свободы соответствует точке на прямой. Другими словами, система двух уравнений с тремя неизвестными задает линию.

В случае оценки дисперсии

$$S_x^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2),$$

сумма

$$s(x_1, x_2, \ldots, x_n) = ((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)$$

имеет при n независимых координатах $n - 1$ степень свободы в силу соотношения

$$s(x_1, x_2, \ldots, x_n) = s(x_1 + C, x_2 + C, \ldots, x_n + C).$$

Действительно, если все переменные x_i увеличить или уменьшить на одно и то же число, то и $\bar{x}$ изменится на это же число, а значит, разности в скобках останутся неизменными.

Если в формулу выборочной оценки дисперсии поставить реальное математическое ожидание случайной величины X, равное a, то сумма

$$s'(x_1, x_2, \ldots, x_n) = ((x_1 - a)^2 + \ldots + (x_n - a)^2),$$

будет иметь n степеней свободы.

На качественном уровне различие таково: в первом случае оценка математического ожидания $\bar{x}$ сдвигается вслед за случайными сдвигами результатов испытаний. Это приводит к тому, что возведенные в квадрат разности $(x_i - \bar{x})^2$ оказываются несколько меньше, чем аналогичные $(x_i - a)^2$. Эта неточность и нейтрализуется множителем.

Таким образом, дисперсию можно оценивать по любой из формул:

$$S_x^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2),$$

и, если реальное математическое ожидание нам из какого-то источника известно,

$$S_x^2 = \frac{1}{n}((x_1 - a)^2 + \ldots + (x_n - a)^2).$$

Глава 3

Проверка статистических гипотез

3.1. Типичные ситуации, требующие использования математической статистики

Основное поле применения математической статистики — это эмпирические исследования. В качестве базового примера мы возьмем реальное исследование, проведенное одной из лабораторий Московского Университета. В реальном эксперименте участвовало около 60 испытуемых, что слишком много для учебного примера. Мы упростим наши данные, чтобы сделать их более наглядными, но при этом полностью сохраним структуру реальных результатов.

Пример 1. Двадцать работников одной из компаний проходили тестирование, оценивающее уровень тревожности. Их стартовые результаты (в условных баллах) были таковы:

$$\{51, 52, 52, 53, 54, 54, 55, 57, 60, 60, 61, 62, 62, 63, 65, 66, 69, 70, 72, 74\}.$$

На основании данных стартового тестирования испытуемые были разбиты на две группы — контрольную и экспериментальную. С этой целью сначала были выделены пары с близкими стартовыми показателями, а затем из каждой пары случано был выбран испытуемый для экспериментальной группы. Второй член пары был отнесен к контрольной группе.

После этого в течение месяца один раз в неделю с экспериментальной группой проводились тренинговые занятия. В заключение тестирование было повторено. Общие результаты эксперимента приведены в таблице. В верхней половине приведены данные экспериментальной группы, в нижней — контрольной. В последнем столбце — выборочные средние значения помещенных в данной строке показателей.

number	1	2	3	4	5	6	7	8	9	10	*M*
start	51	53	54	57	60	61	62	65	70	72	60,5
final	26	42	63	20	36	35	31	13	25	34	31,5
progress	25	21	−9	37	24	26	31	52	45	38	29

number	1	2	3	4	5	6	7	8	9	10	*M*
start	52	52	54	55	60	62	63	66	69	74	60,7
final	67	24	19	34	32	65	64	50	65	59	47,9
progress	−15	28	35	21	28	−3	−1	16	4	15	12,8

Глядя на средние значения результатов экспериментальной группы, можно заметить, что после тренинговых занятий тревожность участников снизилась с 60,5 до 31,5 балла.

Первый важный вопрос, достаточно ли этого изменения, чтобы сказать, что тревожность участников занятий *заметно* снизилась? Не могло ли так случиться, что тревожность участников испытывала некоторые случайные колебания, например, в связи с личными событиями каждого из них, в результате чего и произошло смещение среднего значения тревожности по экспериментальной группе — в этот раз тревожность случайно уменьшилась на 29 баллов, а в другой раз случайно увеличится, скажем, на 50 баллов?

Второй важный вопрос, чем можно удостоверить, что снижение тревожности произошло благодаря проведенным занятиям, а не благодаря какому-то общему фоновому изменению обстановки, окружающей испытуемых, например, первый тестовый замер проводился накануне решающего матча чемпионата мира по футболу, а второй после окончания чемпионата?

На эти вопросы статистика дает такой ответ, какой она может дать — не хуже и не лучше. Как она это делает, будет темой следующих разделов.

3.2. Общий подход

Идея, лежащая в основе методов проверки статистических гипотез легче всего иллюстрируется на примерах азартных игр. Предположим, некто предложил нам простую игру: 100 раз подбросить монету, ставя

по 10 рублей на каждый бросок. Мы выбрали "герб" и в результате проиграли 960 рублей, поскольку 98 раз из 100 выпала "цифра".

В этом случае мы безусловно можем быть уверены, что наш противник ведет нечестную игру, хотя бы он и уверял нас, что такая серия "цифр" может выпасть случайно. Но сколько "цифр" мы готовы принять в качестве случайного результата честной игры? 70 или 80? Математическая статистика выработала общий подход к такого рода ситуациям. Он состоит в следующем.

Выбирается некоторый уровень допустимой ошибки *отвергнуть* гипотезу о *случайном* происхождении полученного результата, *когда он на самом деле случаен* (она называется ошибкой первого рода). В случае описанной выше игры это будет ошибка — назвать подлецом честного человека. Далее выбирается такое число выпадений "цифры", чтобы вероятность получить в результате честного испытания результат такой или худший (для нас) равнялась бы именно выбранному значению.

Например, выберем уровень допустимой ошибки $\alpha = 0{,}05$. В нашем случае количество выпадений "цифры" за 100 бросаний это биномиальная случайная величина $X = B(100, 1/2)$, математическое ожидание которой равно $M = 50$, среднее квадратическое отклонение $\sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot 1/2 \cdot 1/2} = 5$. Чтобы найти граничное значение, воспользуемся нормальным приближением, которое обсуждалось в третьем параграфе седьмой главы.

При $x_\alpha = 1{,}65$

$$\Phi(x_\alpha) \approx 0{,}45$$

Заменяя стандартизованную биномиальную величину близкой к ней нормальной, имеем

$$P\left(\frac{X-50}{5} > 1,65\right) \approx 0{,}05$$

(см. упражнения в конце седьмой главы третьей части). Неравенство $\frac{X-50}{5} > 1{,}65$ можно заменить на эквивалентное $X > 50 + 1{,}65 \cdot 5 = 58{,}25$, поэтому

$$P(X > 58{,}25) \approx 0{,}05.$$

Точная формулировка решения будет такова: на уровне значимости $\alpha = 0{,}05$ мы отвергаем гипотезу о случайном характере результата, если за 100 бросаний монеты выпадает 59 или более "цифр". В приложении к ситуации это означает, что мы отвергаем гипотезу о честной игре, если выпало 59 или больше "цифр".

Скорее всего, причиной этого экстремального результата является нечестность нашего партнера, но возможны и другие объяснения (вмешательство дьявола, неоднородная по каким-то неведомым нам характеристикам монета, гипнотическое вмешательство третьих лиц). Статистика не берется отвечать на вопросы, лежащие вне ее сферы. Она предлагает лишь некоторую общую меру для подобных ситуаций. Если, допустим, мы собираемся заявить нашему противнику, что он нечестный человек, то надо иметь в виду, что вероятность $\alpha = 0{,}05 = 1/20$ означает, что в среднем в одном случае из двадцати при повторении такой игры из 100 бросков симметричной монеты с совершенно честными людьми мы будем получать превышающий граничное значение результат — если в вашей группе больше 20 человек (все они безупречно честные люди), то, сыграв с каждым, вы с большой вероятностью получите хотя бы один результат, который по нашему критерию считается свидетельством нечестности партнера.

С другой стороны, если "цифра" выпала 100 раз подряд, то вероятность 2^{-100} представляется абсолютно ничтожной для того, чтобы поверить в честную случайность. Стало быть, речь может идти только о той или иной границе для принятия решения. Пожалуй, граница на уровне 98 слишком велика, но хорошо ли объявлять нечестным каждого двадцатого?

В каждом случае вопрос решается по-своему. Так, для публикации результата в журналах по психологии считается допустимым уровень значимости $p = 0{,}05$. Это довольно либеральный критерий, поскольку вероятность принять за существенный на самом деле совершенно случайный результат довольно велика.

Но цена ошибки в этом случае считается приемлемой. Действительно важные для психологического знания результаты проверяются и перепроверяются коллегами, поэтому не входят в корпус научно обоснованных результатов благодаря случаю.

Если же в результате статистической проверки решается, например, вопрос об уголовной ответственности, то уровень значимости 0,05 никак не может быть признан удовлетворительным, поскольку цена ошибки очень велика.

Итак, выбор уровня значимости, или, что то же самое, уровня вероятности ошибки первого рода, всегда зависит от внешних по отношению к статистике обстоятельств.

Отметим еще, что ошибкой второго рода называется принятие гипотезы о случайности, когда она на самом деле неверна. Чем больше вероятность ошибки первого рода в данном статистическом исследо-

вании, тем ниже вероятность ошибки второго рода, и наоборот. Подробнее этот вид ошибок в нашем кратком учебнике мы обсуждать не будем.

3.3. t-критерий для одной выборки

В примере 1 были представлены результаты тестирования двух групп работников некоторой компании. Испытуемые первой, экспериментальной группы между двумя тестированиями участвовали в тренинге, направленном на снижение тревожности. Испытуемые второй, контрольной группы между тестированиями тренинг не проходили, а занимались своей обычной работой.

Первый вопрос, на который следует ответить, действительно ли снизилась тревожность у участников экспериментальной группы. У десяти участников снижение тревожности составило соответственно $25, 21, -9, 37, 24, 26, 31, 52, 45, 38$ баллов. Мы хотим убедиться, что этот результат неслучаен.

Как и в разобранном выше случае с монетой, мы начинаем с предположения, что результат случаен и показываем, что это маловероятно.

Рассмотрим ситуацию в общем случае. Пусть наша выборка состоит из n чисел $(x_1, x_2, \ldots, x_n)$, которые являются результатом независимых испытаний одной и той же случайной величины. Случайность результата значит не что иное, как равенство нулю ее математического ожидания.

> **Замечание 1.** Схема рассуждения "от противного" — узловой момент метода проверки статистических гипотез. Если мы хотим доказать, что имет место эффект некоторого воздействия, мы предполагаем, что эффекта нет, т.е. результат его воздействия нулевой, что эквивалентно равенству нулю математического ожидания некоторой случайной величины. Затем, используя это предположение, а также некоторые дополнительные, без которых невозможен расчет вероятностей, мы показываем, что вероятность получить при данных предположениях данный результат ниже назначенного порога (уровня значимости). Это позволяет отвергнуть исходное предположение (об отсутствии эффекта) и утверждать, что эффект имеет место.

Вернемся к выборке $(x_1, x_2, \ldots, x_n)$. Предположим дополнительно, что случайная величина X, результатом испытаний которой является

наша выборка, нормально распределена. Ее математическое ожидание известно (по предположению оно равно нулю), а дисперсия представляет собой неизвестное число D.

Рассмотрим новую случайную величину, заданную формулой, составные части которой нам уже известны.

Напомним, что $\bar{x}$ это среднее арифметическое выборочных значений, а выборочная дисперсия S_x^2 вычисляется по формуле

$$S_x^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2).$$

Составим формулу

$$t = \frac{\bar{x}}{\sqrt{S_x^2}}\sqrt{n}.$$

Если бы вместо выборочной дисперсии S_x^2 в знаменателе стояла реальная дисперсия D случайной величины, испытание которой дает наше единичное наблюдение, то формула

$$\frac{\bar{x}}{\sqrt{D}}\sqrt{n}.$$

задавала бы стандартную нормальную случайную величину, поскольку дисперсия среднего арифметического получается из дисперсии одного наблюдения делением на $\sqrt{n}$, а

$$\frac{\bar{x}}{\sqrt{D}}\sqrt{n} = \frac{\bar{x}}{\sqrt{D}/\sqrt{n}}.$$

Деление на корень из выборочной дисперсии, которая сама случайно отклоняется от реальной, несколько искажает распределение, но тем не менее график распределения случайной величины

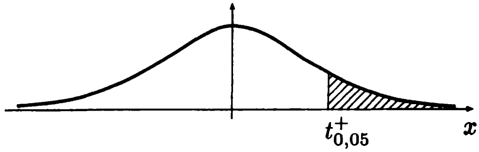


Рис. 3.1

$$t = \frac{\bar{x}}{\sqrt{S_x^2}}\sqrt{n}$$

напоминает график плотности нормального распределения (рис. 3.1). Для разных значений n распределения различны.

Распределение, получаемое по приведенной формуле, если в него входит n слагаемых, называется распределением Стьюдента с $n-1$ степенью свободы. Другое его название t-распределение.

Деление на выборочную дисперсию, хотя и не превращает случайную величину в стандартную нормальную, все же обеспечивает независимость распределения Стьюдента от дисперсии слагаемых. Если исходная случайная величина — это размер горы в метрах и если другая случайная величина измеряет вес бактерий в граммах, то после нормировки каждой из них с помощью ее собственной выборочной дисперсии мы можем употреблять для обеих одно и то же распределение Стьюдента. Это удобство вместе с чрезвычайно широкой распространенностью задачи о средних значениях эмпирических результатов сделало критерий, основанный на распредлении Стьюдента, одним из самых распространенных.

Вернемся к нашему примеру. Мы предположили, что результаты наших испытуемых порождены случайной величной с нулевым математическим ожиданием. Соответствующее распределение Стьюдента также имеет нулевое математическое ожидание. Как и в разобранном случае с монетой, фиксируем $\alpha = 0{,}05$ — вероятность допустимой ошибки первого рода (отвергнуть гипотезу о случайности, когда она верна). На рис. 3.1 точка $t_{0,05}$ — та точка, для которой

$$P(t > t_{0,05}) = 0{,}05.$$

Эта вероятность равна площади заштрихованной фигуры под графиком плотности распределения, расположенной правее $t_{0,05}$.

Если значение t, вычисленное по нашей выборке, окажется больше $t_{0,05}$, то это означает примерно то же, что и 59 “гербов” за 100 бросаний монеты: событие слишком маловероятно, чтобы произойти в отсутствие какого-то систематического смещения среднего значения. Тогда мы отвергаем гипотезу о равенстве нулю математического ожидания случайной величины “изменение уровня тревожности испытуемых экспериментальной группы”.

3.3.1. Практическая реализация

Практическую задачу примера 1 можно решить следующим образом.

1. Заглянуть в таблицу, где приведены граничные значения t-распределения (см. таблицу А). Мы имеем 10 наблюдений, следовательно в строке, соответствующей девяти $(10-1)$ степеням свободы находим

в столбце "Уровень значимости одностороннего[1] критерия 0,05" число 1,83.

2. Вычислить по формуле

$$t = \frac{\bar{x}}{\sqrt{S_x^2}}\sqrt{n}.$$

выборочное значение t (в таких случаях говорят "вычислить t-статистику для данной выборки"). В нашем случае оно равно 5,51.

3. Сравнить граничное значение с вычисленным выборочным. Отвергнуть или принять гипотезу о равенстве нулю среднего значения в зависимости от того, правее или левее граничного значения $t_{0,05}$ расположится полученное выборочное значение t. В нашем случае $t = 5{,}5$ расположено правее $t_{0,05} = 1{,}83$. Это значит, что на уровне значимости $\alpha = 0{,}05$ мы отвергаем гипотезу о равенстве нулю математического ожидания случайной величины "изменение уровня тревожности испытуемых экспериментальной группы".

Таблица А. Распределение Стьюдента. Доверительные границы для t с f степенями свободы

(Сокращенный вариант. Полный вариант в конце книги.)

f	Двухсторонние границы				
	0,1	0,05	0,02	0,01	0,001
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,624	2,977	4,140
15	1,753	2,131	2,602	2,947	4,073
16	1,746	2,120	2,583	2,921	4,015
17	1,740	2,110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,992
	0,05	0,025	0,01	0,005	0,0005
	Односторонние границы				

Упражнение 3.1. Провести аналогичные рассуждения для уровня значимости $\alpha = 0{,}01$.

[1] Об односторонних/двухсторонних критериях речь пойдет в пятом параграфе этой главы.

Для контрольной группы t-статистика равна 2,51.

Упражнение 3.2. Проверить гипотезу о равенстве нулю математического ожидания случайной величины “изменение уровня тревожности испытуемых контрольной группы” для уровней значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$.

Ответ. Для экспериментальной группы гипотеза отвергается на обоих уровнях значимости, для контрольной — только при $\alpha = 0{,}05$, а при $\alpha = 0{,}01$ гипотеза не может быть отвергнута, поскольку соответствующая t-статистика меньше $t_{0{,}01}$.

> **Замечание 2.** Компьютерные статистические пакеты не требуют вычислять показатели прогресса испытуемых. Достаточно задать таблицу начальных и конечных результатов по каждому испытуемому и применить так называемый парный t-критерий. Механизм расчета совпадает с разобранным выше.

Обсуждение. На уровне значимости $\alpha = 0{,}05$ обе группы испытуемых демонстрируют снижение тревожности. Содержательное объяснение этого факта не входит в компетенцию статистики. Можно заметить, что экспериментальная группа демонстрирует несколько большее снижение, поскольку относительно нее гипотеза отвергается на более низком уровне значимости. Однако достаточно ли выражено это преимущество. Не может ли быть так, что случайные колебания t-статистики привели к некоторому преимуществу экспериментальной группы?

3.4. t-критерий для независимых выборок

Для ответа на поставленный вопрос используется вторая модификация t-критерия. Она может использоваться даже в случае, если сравниваемые выборки разного размера. Предположим, мы имеем две выборки, представляющие собой результаты испытаний одной и той же нормальной случайной величины. Первая содержит n наблюдений, а вторая — m наблюдений. Как и прежде, $\bar{x}$ и $\bar{y}$ выборочные средние.

Составим новый вариант t-статистики:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}} \cdot \sqrt{\frac{(n + m - 2)nm}{n + m}}.$$

Здесь $\sum(x_i - \bar{x})^2$ заменяет сумму $(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2$, и аналогично для y. В этом случае мы можем не предполагать, что математическое

ожидание случайной величины, результатами испытаниями которой являются выборки, равно нулю: легко видеть, что математическое ожидание числителя первой дроби всегда равно нулю, если выборки суть реализации одной и той же случайной величины.

Множитель под корнем подобран так, чтобы новая t-статистика имела то же самое распределение Стьюдента (количество степеней свободы равно $(n + m - 2)$).

В нашем примере $n = m = 10$. Вычислим t-статистику, подставив вместо x_i данные экспериментальной группы, а вместо y_i — контрольной. Значение t-статистики равно 2,21.

В таблице распределения Стьюдента с 18 степенями свободы уровню значимости 0,05 соответствует граница $\alpha_{0,05} = 1{,}73$, а уровню значимости 0,01 граница $\alpha_{0,01} = 2{,}55$.

Это значит, что на уровне значимости 0,05 мы можем отвергнуть гипотезу о равенстве средних значений случайных величин, породивших выборки x_i и y_i. Раница между средними значениями снижения тревожности для экспериментальной и контрольной групп слишком велика, чтобы быть результатом случайных вариаций значений одной и той же случайной величины. Однако не так велика, чтобы отвергнуть эту гипотезу с более высокой надежностью на уровне значимости 0,01.

Содержательно эти результаты интерпретировать достаточно легко. По-видимому, какие-то события в среде, в которой живут и работают испытуемые обеих групп, повлияли на средний уровень их тревожности при первом тестировании. Это могли быть глобальные кризисы или вспышки на солнце, повышение цен или падение курса валюты, события непосредственно на предприятии, где проводилось обследование, и даже само обследование, которое могло при первом с ним соприкосновении увеличить уровень тревожности испытуемых.

Однако отличие средних по группам, значимое на уровне 0,05, показывает, что психологический тренинг оказал значительное дополнительное воздействие и привел к более серьезному снижению уровня тревожности в экспериментальной группе.

Заметим, что если бы экспериментаторы не позаботились о тестировании контрольной группы, то могли столкнуться с вполне резонным возражением, что снижение тревожности было вызвано не экспериментальным воздействием, а изменением внешних обстоятельств. Таким образом, в данном примере важно как снижение уровня тревожности экспериментальной группы, так и значимое различие между снижением этого показателя в экспериментальной и контрольной группах.

3.5. Об односторонних и двусторонних критериях

В зависимости от смысла задачи одна и та же статистика может оцениваться с разных позиций.

Проиллюстрируем сказанное примером из области азартных игр. Пусть в игре, о которой шла речь во втором параграфе данной главы, при 100-кратном бросании монеты выпало 60 “гербов”, т.е. мы получили аномальный результат в нашу пользу. К какому выводу должны мы прийти в результате такого испытания? Если нас интересует вопрос о честной игре противника, то, скорее всего, он вне подозрений. Но если нас интересует, является ли симметричной монета, то результат наводит на подозрения. Однако не такие серьезные подозрения.

Дело в том, что вероятность получить при ста бросках результат, отклоняющийся от среднего на девять или больше единиц (безразлично, в какую сторону) равна не 0,05, а 0,1, поскольку складывается из двух симметричных областей $X > 58$ и $X < 42$, каждая из которых дает вероятность 0,05.

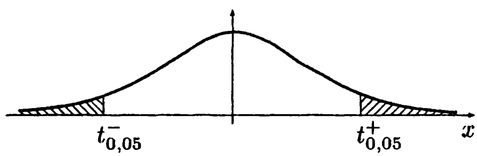


Рис. 3.2

Это в общем случае приводит к тому, что вопрос, который мы задаем природе и на который ждем ответа от статистики, должен формулироваться так, чтобы было ясно, какого рода альтернативу мы предполагаем для основной гипотезы. Если альтернатива явно односторонняя (оценить, нет ли злого умысла в игре нашего противника или достаточно ли выражено позитивное воздействие тренинга), то и критерий может использоваться односторонний (например, основная гипотеза отвергается, если получено аномально большое значение статистики Стьюдента). Если же ставится вопрос, предполагающий обе альтернативы (симметрична ли монета, которую мы бросили 100 раз), то критерий должен браться двусторонний, и гипотеза о симметричности должна отвергаться, если получен аномально большой или аномально низкий результат. На рис. 3.1 показана область отвержения основной гипотезы при одностороннем критерии, на рис. 3.2 — критическая область для двустороннего критерия на графике распределения той же самой случайной величины. В первом случае попадание результата в критическую область означает отвержение основной гипотезы на

во втором попадание в одну из двух симметричных частей критической области означает, что основная гипотеза отвергается на том же уровне значимости 0,05.

3.6. О построении критериев

В следующих главах мы столкнемся с многообразными критериями, предназначенными для решения тех или иных задач. Все они обладают общей логической структурой.

1) Формулируется *основная (или нулевая) гипотеза* H_0. Она предполагает известными некоторые параметры распределения случайной величины, результаты испытаний которой и представлены выборкой или выборками.

2) Формулируется *альтернативная гипотеза* H_1. Как правило, ее принятие может быть интерпретировано в содержательном смысле — как утверждение о связи интересующих нас явлений.

3) Выбирается *статистика* для выбора между H_0 и H_1. Статистика — это та или иная формула (точнее говоря, функция), составленная из элементов выборки или выборок.

4) Для каждого *уровня значимости* α устанавливается *критическая область*, обладающая следующими свойствами:
— содержательно понятно, что попадание результата (а именно вычисленной статистики от имеющихся выборок) в критическую область свидетельствует скорее в пользу H_1, чем H_0;
— вероятность попадания результата (статистики) в критическую область, если гипотеза H_0 истинна, равна α.

Технически операции последнего пункта обеспечиваются тем, что выбранная статистика имеет известное распределение[2]. Далее мы опишем способ построения критической области в случае непрерывной

[2] Напомним, что "статистика имеет распределение" означает следующее: если при выполнении нулевой гипотезы призводить соответствующее число испытаний, затем вычислять по соответствующей формуле значение статистики, то полученные результаты будут случайно варьироваться, т.е. будут представлять собой случайную величину — ее-то распределение и имеется в виду. Например, если бы тренинг не влиял на тревожность, то при многократных испытаниях соответствующая статистика Стьюдента колебалась бы с некоторым разбросом вокруг нулевого значения.

случайной величины. Дискретный случай будет разобран в главе 5 после того, как мы познакомимся с примерами дискретных статистик.

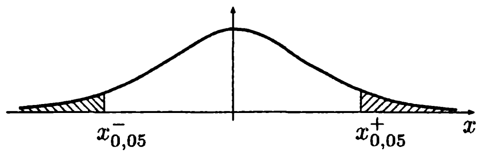


Рис. 3.3

На рис. 3.3 и 3.4 представлены графики плотностей распределения двух известных статистик. Первая имеет симметричную плотность, вторая асимметричную. Обе эти статистики задаются некоторыми формулами, где вместо переменных надлежит подставить выборочные значения. Мы специально не будем уточнять, какие именно статистики здесь представлены, поскольку дальнейшие рассуждения имеют самый общий характер.

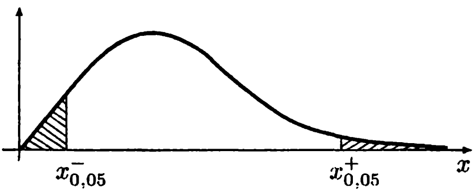


Рис. 3.4

В обоих случаях экстремально большие значения могут быть как угодно велики. Эти их значения в конкретных случаях свидетельствуют скорее в пользу альтернативной гипотезы. Если установлен уровень значимости α, то правая критическая область всегда имеет вид $x > x^+_\alpha$. При этом x^+_α это такое число, что площадь под кривой плотности распределения правее x^+_α равна α, или, что то же самое,

$$\int\limits_{x^+_\alpha}^{\infty} p(x)\,dx = \alpha,$$

где $p(x)$ плотность распределения данной статистики.

Число x^+_α называется *верхним квантилем* распределения данной статистики. Таким образом, находя по таблицам распределения Стьюдента значение t, соответствующее 0,05, мы находили на самом деле верхний квантиль распределения Стьюдента для уровня значимости 0,05 или, как его обычно обозначают, $t^+_{0,05}$.

Нижний квантиль находится точно так же, но левее него попадают экстремально малые значения статистики. Следует обратить внимание

на то, что для статистики, изображенной на рис. 3.3, "малые" означает большие по модулю отрицательные, а для статистики, график которой представлен на рис. 3.4, "малые" это действительно малые близкие к нулю значения.

Тем не менее в обоих случаях для уровня значимости α ищется область с условием, что площадь под графиком плотности распределения также равна α, но только для $x < x_\alpha^-$. Поскольку первый график симметричен, то очевидно $x_\alpha^- = -x_\alpha^+$. Для второго распределения это не так.

Если нам понадобится построить двусторонний критерий (в том случае, если альтернативная гипотеза такова, что в ее пользу говорят как экстремально большие, так и экстремально малые значения) для уровня значимости α, то следует найти квантили $x_{\alpha/2}^+$ и $x_{\alpha/2}^-$ и объединить критические области.

Для симметричных распределений в силу сказанного выше имеет смысл говорить о двустороннем квантиле x_α для статистики St, который удовлетворяет равенству

$$P(St > x_\alpha) + P(St < -x_\alpha) = \alpha.$$

Последнее равенство компактнее можно записать в эквивалентном виде:

$$P(|St| < x_\alpha)) = 1 - \alpha,$$

поскольку события

$$(St < -x_\alpha),\ (|St| < x_\alpha),\ (St > x_\alpha)$$

представляют собой полную систему.

Упражнение 3.3. Какой из верхних квантилей некоторой произвольной статистики будет лежать на числовой оси правее, $x_{0,02}^+$ или $x_{0,01}^+$?

Какой из нижних квантилей окажется правее, $x_{0,02}^-$ или $x_{0,01}^-$?

Глава 4

Распределения хи-квадрат и Стьюдента

4.1. Доверительный интервал для среднего значения

В предыдущей главе мы познакомились с двумя вариантами использования распределения Стьюдента. Для более точного разговора об этом распределении нам понадобится второе важное распределение — так называемое распределение χ^2.

Если $\xi_1, \ldots, \xi_n$ – независимые стандартные случайные величины $N(0,1)$, то

$$\chi_n^2 \stackrel{\text{def}}{=} \xi_1^2 + \xi_2^2 + \cdots + \xi_n^2.$$

Распределение случайной величины χ_n^2 называется распределением *хи-квадрат с n степенями свободы*. Заметим, что случайная величина χ_n^2 принимает только положительные значения при любом n.

Отметим также, что

$$M\chi_n^2 = n, \qquad D\chi_n^2 = 2n.$$

Если дополнительно ξ_0 также имеет распределение $N(0,1)$, то распределение случайной величины

$$t_n \stackrel{\text{def}}{=} \xi_0 \Bigg/ \sqrt{\frac{\chi_n^2}{n}},$$

и называется t-распределением или распределением *Стьюдента с n степенями свободы.*

Поскольку квантили стандартного нормального закона $N(0,1)$ обладают свойством симметрии относительно нуля, то аналогичное свойство симметрии справедливо и для квантилей распределения Стьюдента. В частности, математическое ожидание $Mt_n = 0$. Пусть $t_\alpha(n)$ обозначает двусторонний квантиль распределения Стьюдента с n степенями свободы, т.е.

$$P(t_n \le -t_\alpha(n)) = \alpha/2, \quad P(t_n \ge t_\alpha(n)) = \alpha/2,$$

$$P(-t_\alpha(n) < t_n < t_\alpha(n)) = 1 - \alpha.$$

Из теоретико-вероятностного закона больших чисел следует, что при $n \to \infty$ распределение $t_n \to N(0,1)$. При этом для любого $n \ge 2$ квантиль $t_\alpha(n) > x'_\alpha$ и

$$\lim_{n\to\infty} t_\alpha(n) = x'_\alpha,$$

где x'_α – двусторонний квантиль распределения $N(0,1)$

По наблюдениям $\mathbf{x} = (x_1, x_2, \dots, x_n)$, как мы знаем, вычисляются выборочное среднее $\bar{x}$ и выборочная дисперсия S_x^2:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \qquad S_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2.$$

Символом s обозначается выборочное среднеквадратическое отклонение:

$$s = \sqrt{S_x^2}.$$

Данные величины (статистики) называются *точечными оценками* соответствующих теоретических параметров a, D и σ. Это значит, что наиболее правдоподобными значениями реальных математического ожидания, дисперсии и среднеквадратического отклонения случайной величины, которую представляет данная выборка, как раз и являются полученные нами оценки.

Пример 1. Предположим, вы измерили рост первых 100 студентов, входящих в ваше учебное заведение. Средний рост этой сотни и есть наиболее вероятное значение среднего роста всей совокупности студентов данного вуза.

Однако понятно, что наша оценка никогда не попадает в точное значение параметра, т.е. реальный параметр отклоняется от оценки. Для того чтобы можно было судить о диапазоне возможного отклонения, вводится понятие *доверительного интервала.*

Упражнение 4.1. Проверить, что

$$\frac{x_1 + \cdots + x_n}{n} - a = \frac{(x_1 - a) + \cdots + (x_n - a)}{n}.$$

Если реальное математическое ожидание некоторой нормальной случайной величины равно a, то

$$\frac{(\bar{x} - a)\sqrt{n}}{s}$$

имеет распределение Стьюдента с n степенями свободы (см. определение распределения Стьюдента и упражнение 1).

В таком случае для этой случайной величины вероятность попадания в интервал между (двусторонними) квантилями выражается формулой

$$P\left(-t_\alpha(n-1) < \frac{(\bar{x} - a)\sqrt{n}}{s} < t_\alpha(n-1)\right) = 1 - \alpha.$$

Неравенство в скобках можно преобразовать:

$$-t_\alpha(n-1) < \frac{(\bar{x} - a)\sqrt{n}}{s} < t_\alpha(n-1)$$

эквивалентно неравенству

$$-\frac{s \cdot t_\alpha(n-1)}{\sqrt{n}} < (\bar{x} - a) < \frac{s \cdot t_\alpha(n-1)}{\sqrt{n}},$$

а это неравенство алгебраически эквивалентно следующему:

$$\bar{x} - \frac{s \cdot t_\alpha(n-1)}{\sqrt{n}} < a < \bar{x} + \frac{s \cdot t_\alpha(n-1)}{\sqrt{n}}.$$

Возвращаясь к вероятности, заменяем выражение в скобках на последний вариант неравенства:

$$P\left(\bar{x} - \frac{s \cdot t_\alpha(n-1)}{\sqrt{n}} < a < \bar{x} + \frac{s \cdot t_\alpha(n-1)}{\sqrt{n}}\right) = 1 - \alpha.$$

Величину

$$\epsilon_\alpha = \frac{s \cdot t_\alpha(n-1)}{\sqrt{n}}$$

называют радиусом доверительного интервала. Окончательно

$$P(\bar{x} - \epsilon_\alpha < a < \bar{x} + \epsilon_\alpha) = 1 - \alpha.$$

Поскольку реальное математическое ожидание a не является случайной величиной, то последнюю формулу, вообще говоря, нельзя интерпретировать как вероятность попадания реального математического ожидания в интервал $[\bar{x} - \epsilon_\alpha;\ \bar{x} + \epsilon_\alpha]$, однако, сделав эту оговорку, мы можем иногда придавать доверительному интервалу вероятностный смысл, имея в виду скорее *обыденное* понятие вероятности, чем строго теоретическое.

Строгая же формулировка такова: *при коэффициенте доверия* $1-\alpha$ *параметр* a *заключен в доверительный интервал с границами:*

$$a_\alpha^- = \bar{x} - \epsilon_\alpha, \quad a_\alpha^+ = \bar{x} + \epsilon_\alpha.$$

Значения a, лежащие внутри интервала, со степенью доверия $1-\alpha$ согласуются с наблюдениями, представленными выборкой. Значения вне интервала — не согласуются.

Мы рассматривали задачу проверки статистической гипотезы о равенстве нулю математического ожидания некоторой случайной величины. Построение доверительного интервала решает целую серию подобных задач — фактически каждое a проверяется по одной и той же данной выборке и в результате проверки принимается в качестве возможного или отвергается.

В отличие от точечной оценки математического ожидания $\bar{x}$, доверительный интервал называют *интервальной оценкой*.

Упражнение 4.2. Какой интервал шире, соответствующий коэффициенту доверия 0,9 или 0,98? Каков радиус доверительного интервала с коэффициентом доверия 1?

Ответ. Шире интервал, соотвтетствующий $1 - \alpha = 0{,}98$. Коэффициенту 1 соответствует бесконечный доверительный интервал. Содержательно это означает, что при данной выборке стопроцентно мы можем гарантировать лишь попадание реального математического ожидания куда-нибудь на всю числовую прямую.

Замечание 1. Если параметр *σ известен*, то задача построения доверительного интервала для *неизвестного* параметра a упрощается: При реальном математическом ожидании

нормальной случайной величины, равном a, и среднеквадратическом отклонении, равном σ,

$$\frac{(\bar{x} - a)\sqrt{n}}{\sigma}$$

имеет нормальное распределение $N(0, 1)$, поэтому радиус доверительного интервала считается с помощью квантилей нормального распределения.

Для стандартного нормального распределения двусторонний квантиль x'_α это число, удовлетворяющее равенству

$$\Phi(x'_\alpha) = \alpha/2.$$

Проведя совершенно аналогичные выкладки, можем убедиться, что радиус доверительного интервала в этом случае находится по формуле

$$\epsilon_\alpha = \frac{\sigma \cdot x'_\alpha}{\sqrt{n}}.$$

4.2. Критерий согласия χ^2 (хи-квадрат)

В разделе о проверке биномиальных гипотез мы проверяли гипотезу о равенстве неизвестной вероятности некоторому числу. Подчеркнем, что речь шла об уточнении значения *одного* параметра — вероятности. Иной характер имеет ситуация, когда требуется проверить гипотезу о равенстве определенным значениям *нескольких* вероятностей (иначе говоря, о законе распределения в целом). В таких случаях применяются так называемые *критерии согласия*. Мы рассмотрим один из них, связанный с распределением χ^2.

Пусть в результате некоторого испытания может произойти одно из k событий $A_1, A_2, \cdots, A_k$. Нулевая гипотеза H_0 имеет следующий вид:

$$p(A_1) = p_1, \quad p(A_2) = p_2, \quad \ldots, \quad p(A_k) = p_k,$$

где $p_1, p_2, \cdots, p_k$ — некоторые положительные числа, сумма которых равна 1. Альтернативной гипотезой является невыполнение хотя бы одного из этих равенств.

Исходными данными для проверки гипотезы H_0 являются результаты n независимых испытаний. Пусть в результате них

событие A_1 произошло m_1 раз,
событие A_2 произошло m_2 раз,
...
событие A_k произошло m_k раз.

Очевидно, что

$$m_1 + m_2 + \ldots + m_k = n.$$

Величина

$$\frac{(m_1 - np_1)^2}{np_1} + \frac{(m_2 - np_2)^2}{np_2} + \cdots + \frac{(m_k - np_k)^2}{np_k}$$

имеет распределение близкое к χ^2, если n достаточно велико. Таким образом, для проверки гипотезы надо вычислить величину

$$\chi^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i}.$$

Эта величина показывает, насколько экспериментальные значения m_i расходятся с теоретически наиболее вероятными значениями np_i.

Далее проверка гипотезы осуществляется уже привычным образом.

Число степеней свободы f на единицу меньше, чем количество возможных исходов $f = k - 1$.

Выбирается уровень значимости (например, $\alpha = 0{,}05$ либо $\alpha = 0{,}01$). после чего находится критическое значение $\chi^+_\alpha(f)$ (зависящее от α и f) по таблицам, содержащим квантили распределения χ^2:

Таблица Б. Распределение χ^2.
Доверительные границы для χ^2 с f степенями свободы
(Сокращенный вариант. Полный вариант в конце книги.)

f	$\alpha = 0{,}2$	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
1	1,642	2,706	3,841	6,635
2	3,219	4,605	5,991	9,210
3	4,642	6,251	7,815	11,345
4	5,989	7,779	9,488	13,277
5	7,289	9,236	11,070	15,086
6	8,558	10,645	12,592	16,812
7	9,803	12,017	14,067	18,475
8	11,030	13,362	15,507	20,090
9	12,242	14,684	16,919	21,666
10	13,442	15,987	18,307	23,209

Если

$$\chi^2 > \chi_\alpha^+(f),$$

то гипотеза H_0 отвергается, в противном случае — принимается.

Мы упоминали о том, что описанный критерий применим лишь при достаточно больших значениях n. Точнее, требование формулируется так: все величины np_i должны быть достаточно велики. Уверенно критерий можно применять при

$$np_i > 10, \quad i = 1, 2, \ldots, r.$$

Пример 2. При 4040 бросаниях монеты французский естествоиспытатель Бюффон получил 2048 выпадений герба и 1992 выпадений цифры. На уровне значимости $\alpha = 0{,}05$ проверим гипотезу о том, что монета была правильной.

Решение. Здесь в результате испытания может произойти одно из двух событий — выпадение герба либо выпадение цифры. Поэтому имеем:

$A_1 = \{$выпадение герба$\}$, $A_2 = \{$выпадение цифры$\}$,

$n = 4040$, $m_1 = 2048$, $m_2 = 1992$.

Нулевая гипотеза —

$$H_0: \quad p(A_1) = p(A_2) = \frac{1}{2},$$

то есть

$$p_1 = p_2 = \frac{1}{2}.$$

Вычислим величину χ^2. Имеем

$$\chi^2 = \frac{(m_1 - np_1)^2}{np_1} + \frac{(m_2 - np_2)^2}{np_2} =$$

$$= \frac{(2048 - 2020)^2}{2020} + \frac{(1992 - 2020)^2}{2020} \approx 0{,}776.$$

Число степеней свободы f в данном случае равно $2 - 1 = 1$. По известным значениям $\alpha = 0{,}05$, $f = 1$ находим в таблице значение

$$\chi_{0,05}^+(1) = 3{,}8.$$

Так как

$$\chi^2 < \chi_{0,05}^+(1),$$

то нулевая гипотеза принимается — монета была правильной.

4.3. Проверка соответствия эмпирической функции распределения нормальному закону

Наиболее часто критерий χ^2 применяется для проверки гипотезы о нормальном распределении случайной величины, в результате испытаний которой была получена данная выборка.

Пример 3. Пусть в результате измерений роста студентов получена выборка, содержащая 500 наблюдений, заключенных между 150 и 190 см. Мы не будем приводить всю выборку, а предположим, что данные введены в компьютер и мы можем производить все необходимые вычисления.

Прежде всего следует вычислить выборочные среднее и среднеквадратическое отклонение. Допустим, для нашей выборки результаты таковы $\bar{x} = 170$ см и $s = 10$.

Далее мы должны устанавить интервалы разбиения. Мы выбираем симметричную относительно $\bar{x}$ систему границ 155, 165, 175, 185, разбивающую числовую прямую на пять интервалов, два из которых неограничены с одной из сторон. Хотя реальные результаты измерения роста не могут быть ни отрицательными, ни очень большими положительными, рассматривая соответствие нашего выборочного распределения нормальному закону, мы можем не беспокоиться об этих крайностях, поскольку, например, вероятность получить отрицательное значение нормально распределенной случайной величины с параметрами, расчитанными по нашей выборке, неотличима от нуля.

Теперь мы подсчитываем количество выборочных значений, попадающих в наши интервалы, и записываем результаты в таблицу (3-й столбец таблицы расчетов). Имеется 28 результатов измерений меньших 155, 130, лежащих в интервале от 155 до 165, и т.д. (см. таблицу расчетов).

Далее нам предстоит вычислить теоретические вероятности попадания в эти же интервалы значений нормально распределенной случайной величины с параметрами, равными расчитанным по нашей выборке ($a = 170$ и $s = 10$). Для этого мы воспользуемся значениями функции $\Phi(x)$. При $s = 10$, $155 = 170 - 1{,}5s$ и для стандартной нормальной случайной величины ξ

$$P(\xi < -1{,}5) = P(\xi > 1{,}5) = 1 - \Phi(1{,}5) = 0{,}067\,,$$

поэтому в крайние ячейки четвертого столбца таблицы расчетов записываются вероятности 0,067.

Далее $165 = 170 - 0{,}5s$, поэтому

$$P(-1{,}5 < \xi < -0{,}5) = P(0{,}5 < \xi < 1{,}5) = \Phi(1{,}5) - \Phi(0{,}5) = 0{,}241\ .$$

В клетки четвертого столбца таблицы, соответствующие интервалам (155, 165) и (175, 185), записываются теоретические вероятности 0,241.

На долю центрального интервала остается вероятность 0,384.

В последнем столбце записывается ожидаемое число попаданий в данный интервал, т.е. значение вероятности, помещенной в четвертом столбце, умноженное на количество наблюдений, равное 500 (т.е. np_i).

Таблица расчетов

Номер интервала	Интервал	Число значений, попавших в интервал	Теоретическая вероятность для нормального распределения	Число попаданий, предсказанных нормальным распределением
1	$(-\infty, 155)$	28	0,067	33
2	(155, 165)	130	0,241	121
3	(165, 175)	194	0,384	192
4	(175, 185)	110	0,241	121
5	$(185, \infty)$	38	0,067	33

Теперь для каждой строки необходимо подсчитать значение выражения

$$\frac{(n_i - np_i)^2}{np_i}$$

и сложить пять полученных чисел. Расчет дает:

$$\frac{5^2}{33} + \frac{9^2}{121} + \frac{2^2}{192} + \frac{11^2}{121} + \frac{5^2}{33} =$$

$$= \frac{25}{33} + \frac{81}{121} + \frac{4}{192} + \frac{121}{121} + \frac{25}{33}.$$

Производя операции, получаем $= 0{,}76 + 0{,}67 + 0{,}02 + 1 + 0{,}76 = 3{,}21$.

Теперь важное замечание. В данном случае количество степеней свободы равно $5 - 3 = 2$. Объяснение, которое было дано в конце второй главы по поводу числа степеней свободы при оценке дисперсии, распространяется на этот случай.

Наша статистика χ^2 связана тремя соотношениями:
— во-первых, имеется соотношение, "сумма наблюдений по интервалам равна общему количеству наблюдений";
— во-вторых, при вычислении ожидаемых значений попаданий в интервалы использовалось выборочное среднее $\bar{x}$;
— в-третьих, при вычислении ожидаемых значений попаданий в интервалы использовалось выборочное значение среднеквадратического отклонения s.

В общем случае чтобы получить число степеней свободы для нашего критерия χ^2, надо из числа интервалов n вычесть 3 — количество соотношений. Для нашего примера, как уже отмечалось, это $5-3=2$.

Пятипроцентный односторонний квантиль $\chi^+_{0,05}(2)$ равен 5,99. Расчитанное по выборке значение, равное 3,21, меньше граничного, следовательно, на уровне значимости 0,05 принимается гипотеза H_0 о соответствии выборки нормальному распределению.

Глава 5

Непараметрические аналоги t-критерия

В разделе "О построении критериев" третьей главы мы описали общую логическую структуру критериев, предназначенных для проверки гипотез статистическими методами.

Мы рассмотрели реализацию этого подхода для нескольких задач, связанных с непрерывными распределениями. Обратим внимание на то, что во всех использованных методах непременно требовалось нормальное распределение случайных величин, испытания которых порождают анализируемые выборки. Хотя нормальное распределение всегда появляется, когда имеет место суммирование воздействий большого числа независимых факторов, но все же ситуация не так проста. Например, если нормально распределены веса капель, падающих с крыши, то их диаметры уже не будут распределены нормально.

Для того чтобы отмести самые грубые несоответствия выборок нормальному распределению, можно использовать критерий χ^2, описанный в предыдущей главе. Если гипотеза о нормальности распределения отвергается на уровне значимости 0,05, то возможность использования t-статистики оказывается под вопросом.

Для работы в этом и подобных случаях разработаны так называемые *непараметрические методы*

В рассмотренной в третьей главе типичной задаче от статистических методов требовалось ответить на два вопроса:

1) можно ли утверждать, что выборка, характеризующая прогресс испытуемых при некотором воздействии, имеет существенно отличное от нуля среднее значение?

2) можно ли утверждать, что среднее значение прогресса экспериментальной выборки существенно выше, чем аналогичное среднее контрольной группы?

В случае, если по изложенным выше обстоятельствам вызывает сомнение возможность применения t-критерия, для решения первой задачи можно применить критерий знаков или критерий знаковых рангов, а для второй — критерий Манна—Уитни.

Все три эти критерия принадлежат к семейству непараметрических, что означает, что их требования к распределению порождающих выборки случайных величин минимальны.

5.1. Критерий знаков и критерий знаковых рангов Вилкоксона

Экспериментальная группа из десяти испытуемых в эксперименте, описанном в третьей главе, показала некоторое снижение тревожности, которое выражалось следущими баллами: 25, 21, −9, 37, 24, 26, 31, 52, 45, 38 со среднем значением 12,8.

Построим критерий знаков. Рассмотрим гипотезу H_0, утверждающую, что для всякого испытуемого улучшение или ухудшение показателя тревожности равновероятно. Смысл критерия знаков состоит в том, что гипотеза H_0 будет отвергнута, если количество улучшивших показатели тревожности испытуемых будет аномально велико. Если вероятность улучшить/ухудшить показатель тревожности равна 0,5, то количество его улучшивших среди десяти испытуемых будет распределено биномиально (точно так же, как количество "гербов" при десятикратном бросании монеты).

Наша задача научиться находить квантили для биномиального распределения. Пусть установлен уровень значимости 0,05. Алгоритм поиска верхнего одностороннего квантиля таков:

1) последовательно записываем вероятности получить 10, 9, ... "гербов" при десяти бросаниях монеты

$$p_{10} = C_{10}^0/2^{10} = 1/1024,$$

$$p_9 = C_{10}^1/2^{10} = 10/1024,$$

$$p_8 = C_{10}^2/2^{10} = 45/1024;$$

2) вычисляем суммы:

$$\begin{aligned}
&p_{10} = 0{,}00097\\
&p_{10} + p_9 = 0{,}01074\\
&p_{10} + p_9 + p_8 = 0{,}05469\\
&\ldots
\end{aligned}$$

3) берем последнюю строку, содержащую сумму меньшую или равную установленному уровню значимости (в нашем случае для $\alpha = 0{,}05$ это вторая строка).

Верхним односторонним квантилем биномиального распределения для уровня значимости $\alpha = 0{,}05$ является число выпадений "герба", равное 9. Если в результате испытания получено девять или десять "гербов", то гипотеза о равной вероятности "герба" и "цифры" отвергается. Если в нашем эксперименте получено, что у девяти из десяти испытуемых тревожность снизилась, то гипотеза о равной вероятности снижения или повышения тревожности между стартовым и финальным тестированиями отвергается на уровне значимости 0,05.

Упражнение 5.1. Отвергается ли гипотеза H_0 на уровне значимости $\alpha = 0{,}01$?

Ответ: Последняя строка, оканчивающаяся суммой, меньшей, чем 0,01, — это первая строка. Следовательно, верхний односторонний квантиль для $\alpha = 0{,}01$ равен 10, т.е. гипотеза отвергается, только если тревожность понизилась у всех испытуемых.

Упражнение 5.2. Рассмотрим контрольную группу с результатами 28, −15, 35, 21, 28, −3, −1, 16, 4, 15.

Будет ли для этой выборки отвергнута на уровне значимости 0,05 гипотеза о равной вероятности снижения и повышения тревожности?

Ответ: не будет.

Упражнение 5.3. Если для произвольного статистического критерия гипотеза H_0 отвергнута на уровне значимости $\alpha = 0{,}01$, следует ли из этого, что она отвергается также и на уровне значимости $\alpha = 0{,}05$?

5.1.1. Критерий знаковых рангов

Критерий знаков работает таким образом, что наша выборка экспериментальной группы (25, 21, −9, 37, 24, 26, 31, 52, 45, 38) для него не отличается от выборки (1, 1, −1, 1, 1, 1, 1, 1, 1, 1), поскольку рассматриваются только знаки входящих в выборку чисел.

Однако первая выборка обладает еще одним свойством, свидетельствующим в пользу того, что смещение в сторону положительных значений неслучайно. Можно обратить внимание на то, что единственное отрицательное число в выборке оказывается к тому же самым маленьким по модулю. Здравый смысл подсказывает, что такая картина наблюдается в том случае, если случайная величина колеблется вокруг положительного среднего значения, редко и ненамного "переходя" в отрицательную область.

Для того чтобы критерий уловил эту тенденцию, он должен учитывать абсолютные величины членов выборки. Таким качеством обладает критерий знаковых рангов Вилкоксона[1].

Таблица В. Распределение Вилкоксона.
Нижние граничные значения.
(Сокращенный вариант. Полный вариант в конце книги.)

n	односторонний 1% двусторонний 2%	односторонний 2,5% двусторонний 5%	односторонний 5% двусторонний 10%
5			1
6		1	2
7	0	2	4
8	2	4	6
9	3	6	8
10	5	8	11

Для того чтобы применить критерий Вилкоксона к данной выборке, надо

1) установить уровень значимости (предположим $\alpha = 0{,}05$) и найти соответствующий (*нижний*) квантиль распределения Вилкоксона (см. таблицу В и полную таблицу в конце книги); для $n = 10$ односторонний пятипроцентный квантиль равен 11;

2) расположить все члены выборки в порядке возрастания абсолютной величины, подписать под ними их ранги; в нашем примере это будет следующий порядок и ранги:

[1] Иногда в литературе этот критерий носит имя Манна—Уитни, а рассмотренный ниже в нашей книге критерий Манна—Уитни, напротив, называется критерием Вилкоксона. Без всякой путаницы рассматриваемый здесь критерий можно называть непараметрическим парным критерием знаковых рангов (для каждого испытуемого рассматриваются пары значений — до и после воздействия), а разбираемый в следующем параграфе критерий — непараметрическим критерием сравнения средних для независимых выборок.

−9	21	24	25	26	26	31	37	38	45
1	2	3	4	5	6	7	8	9	10

3) вычислить статистику Вилкоксона, для чего подсчитать сумму рангов, приписанных отрицательным членам выборки (в нашем случае 1);
4) сравнить полученную статистику с найденным ранее квантилем.

Если полученная сумма рангов меньше значения квантиля, то гипотеза H_0 о случайном характере смещения выборки в сторону положительных значений должна быть отвергнута (в нашем случае $1 < 11$, гипотеза H_0 отвергается).

Замечание 1. Мы использовали односторонний критерий, проверяя гипотезу H_0 о равенстве нулю математического ожидания против гипотезы H_1: “математическое ожидание больше нуля”. Мы вычисляли сумму рангов отрицательных членов выборки и отвергли гипотезу H_0, поскольку эта сумма рангов оказалась меньше соответствующего нижнего квантиля.

Полностью эквивалентен этому другой вариант: подсчитать сумму рангов положительных членов выборки и отвергнуть H_0, если эта сумма превосходит верхний односторонний квантиль того же уровня значимости. Первый вариант предпочтительнее только в силу простоты вычисления меньшей суммы.

Пример 1. Проверить гипотезу о равенстве нулю математического ожидания для выборки

$$1, 2, -5, -6, 7, 8, 8, 9, 10.$$

Сумма рангов отрицательных членов выборки $3 + 4 = 7$.

В таблице распределения Вилкоксона находим для $n = 9$ пятипроцентный квантиль $w^-_{0,05} = 8$, а $w^-_{0,025} = 6$.

Вывод: H_0 отвергается на уровне значимости 0,05, но не может быть отвергнута на уровне значимости 0,025.

Замечание 2. В нашей выборке имеются повторяющиеся значения, равные 8; они занимают седьмое и восьмое места в выборке. При подсчете суммы положительных рангов, этим числам следует приписать равный ранг так, чтобы не изменить общую сумму рангов: в нашем случае это ранги $(7 + 8)/2 = 7{,}5$.

5.2. Критерий Манна—Уитни для независимых выборок

В третьей главе мы решали задачу сравнения средних значений двух выборок. В общем случае выборки $x_1, \dots, x_n$ и выборки $y_1, \dots, y_m$ n может быть не равно m.

Мы использовали t-критерий, который требует, чтобы выборки были реализациями нормально распределенных случайных величин, имеющих одинаковую дисперсию. Это значит, что графики плотности распределения случайных величин X и Y представляют собой подобные кривые нормального распределения, отличающиеся разве что сдвигом. При этих условиях с помощью t-критерия проверяется наличие существенного сдвига или его отсутствие.

Если условие нормальности не гарантируется, к тем же выборкам можно приментить критерий Манна—Уитни. Требование подобия распределений сохраняется, но эти распределения не обязаны быть нормальными.

Напомним наши результаты:

$$\begin{array}{lcccccccccc} x: & 25 & 21 & -9 & 37 & 24 & 26 & 31 & 52 & 45 & 38 \\ y: & 28 & -15 & 35 & 21 & 28 & -3 & -1 & 16 & 4 & 15 \end{array}$$

(вверху экспериментальная группа, внизу контрольная).

Запишем члены обеих выборок в порядке возрастания, выделяя курсивом экспериментальную и жирным шрифтом контрольную группы.

-15, *-9*, **-3**, **-1**, **4**, **15**, **16**, **21**, *21, 24, 25, 26,* **28**, **28**, *31,* **35**, *37, 38, 45, 52*

Если бы случайные величины X и Y имели бы совершенно одинаковое распределение, то "жирные" и "курсивные" числа перемешивались бы более или менее равномерно. Если же экспериментальная группа имеет систематически больший балл, то "курсивные" числа должны оказаться в основном правее "жирных" — как оно и оказывается в нашем случае. Пока неясно только, достаточно ли много членов экспериментальной выборки оказалось правее членов контрольной выборки, чтобы мы могли утверждать наличие систематического сдвига.

Для проверки гипотез мы должны вычислить статистику Манна—Уитни и воспользоваться далее таблицами распределения (Таблица Г в конце книги), в которых для различных уровней значимости даны граничные значения этой статистики.

Подробнее:

1) подсчитываем для каждого "жирного" числа, сколько "курсивных" чисел расположены левее него. Если "жирное" число равно "курсивному", то прибавляем 0,5 (в нашем случае левее "жирного" числа 21 имеется одно строго меньшее "курсивное" и одно "курсивное" ему равно, поэтому для числа 21 результат равен 1,5); в итоге имеем последовательность результатов: 0, 1, 1, 1, 1, 1, 1,5, 5, 5, 6. Складываем эти числа, получаем $0 + 1 + 1 + 1 + 1 + 1 + 1{,}5 + 5 + 5 + 6 = 22{,}5$.

2) смотрим на выбранном уровне значимости (предположим, 0,05) нижний квантиль распределения Манна—Уитни для выборок объема (10, 10) — он равен 27;

3) поскольку левее "жирной" выборки расположено аномально мало членов "курсивной" выборки (22,5 при 5-процентном квантиле 27), мы отвергаем гипотезу H_0 о случайном характере сдвига среднего значения экспериментальной выборки относительно среднего значения выборки контрольной группы. Тем самым утверждается, что экспериментальная группа продемонстрировала на пятипроцентном уровне значимо большее снижение тревожности.

К полностью эквивалентному результату приведет и следующий путь вычисления:

1) подсчитываем для каждого "курсивного" числа, сколько "жирных" чисел расположены левее него (в нашем случае это следующие 10 чисел: $1, 6{,}5, 7, 7, 7, 9, 10, 10, 10, 10$), и затем складываем эти числа ($1 + 6{,}5 + 7 + 7 + 7 + 9 + 10 + 10 + 10 + 10 = 77{,}5$).

2) смотрим на выбранном уровне значимости верхний квантиль распределения Манна—Уитни для числа наблюдений (10, 10);

3) поскольку левее "курсивной" выборки расположено аномально много членов "жирной" выборки (77,5 при 5-процентном верхнем квантиле 73), мы отвергаем гипотезу H_0 о случайном характере сдвига среднего значения экспериментальной выборки относительно среднего значения выборки контрольной группы.

Тождественность вывода объясняется тем, что распределение Манна—Уитни симметрично. Для выборок (10, 10) число всевозможных пар (x, y), составленных из элементов двух выборок равно 10×10. Всегда в части из них $x > y$ (все они были учтены при первом способе расчета), а в оставшихся $x < y$ (они сосчитаны при втором способе). В сумме количество тех и других дает 100. Если математические ожидания случайных величин X и Y совпадают, то получение аномально большого результата имеет ту же вероятность, что и получение аномально маленького результата. Верхний и нижний квантили расположены

симметрично относительно среднего значения 50, поэтому два способа расчета дают одинаковые результаты.

5.3. Некоторые замечания о статистической работе

Читатель приобрел уже некоторый опыт работы с различными статистическими критериями. Теперь представляется своевременным обсудить некоторые важные вопросы. Повторим сначала уже известное.

При всем различии построения критериев читателю следует осознать то общее, что их объединяет. Сделав это усилие понимания, мы легко поймем дальнейший материал и получим возможность дальнейшего продвижения в статистических методах.

Любой статистический критерий строится примерно в следующей последовательности:

— Формулируется пара альтернатив H_0 и H_1, на различение которых направлен критерий.

— Задается *статистика*, т.е. формула, составленная из случайных величин, в которую при расчете будут подставлены полученные результаты (выборка).

— Расчитывается распределение *статистики* при условии истинности гипотезы H_0.

Критерий, основанный на данной статистике, оказывается полезным, если аномальные значения *статистики* (очень большие, очень маленькие или те и другие вместе) указывают на предпочтительную истинность гипотезы H_1. В этом случае составляются таблицы распределения статистики, которые помогают найти квантили для уровня значимости, который пользователь считает разумным для своих целей, — т.е. выделяются границы области аномальных значений, маловероятных при условии истинности гипотезы H_0.

Если статистика, вычисленная на значениях из выборки, попадает в область аномальных значений, это считается убедительным свидетельством в пользу гипотезы H_1.

Некоторые распределения оказываются полезными в большом количестве разных критериев. К таким относятся уже упоминавшиеся распределения Стьюдента и χ^2. Другие расчитываются специально для одной задачи. Две задачи, которые в случае нормального распределения решаются с помощью одного и того же распределения Стьюден-

та, решаются с помощью двух разных распределений (Вилкоксона и Манна—Уитни), если отказаться от требования нормальности.

В заключение несколько слов о современных компьютерных средствах, которые приходят на смену статистическим таблицам. Если вы "поручите" компьютеру считать нужную вам *статистику* и произвести выбор между гипотезами, то вместо ответа "на уровне значимости 0,05 гипотеза H_0 отвергается" компьютер выведет на экран ответ "$p = 0{,}035$". Это значит, что для любого уровня значимости, большего или равного 0,035, гипотеза H_0 будет отвергнута, а для любого уровня значимости меньшего, чем 0,035, гипотеза H_0 будет принята. Тем самым компьютерные статистические пакеты дают возможность выстраивать непрерывную шкалу надежности статистического вывода.

Глава 6

Точечные оценки и доверительные интервалы для непараметрических аналогов t-критерия

6.1. Распределение Вилкоксона

Знаково-ранговое распределение Вилкоксона можно определить следующим образом. Рассмотрим случайную величину W_n, заданную суммой

$$W_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} w_i,$$

где слагаемые w_i, $i = 1, 2, \ldots, n$ являются независимыми случайными величинами с распределениями

$$w_i(x) \stackrel{\text{def}}{=} \begin{cases} 0, & \text{с вероятностью } 1/2 \\ i, & \text{с вероятностью } 1/2 \end{cases}.$$

Эту случайную величину, имеющую распределение Вилкоксона, можно реализовать следующим физическим процессом.

Пример 1. В урне имеется n жетонов, занумерованных числами от 1 до n. Номер каждого жетона на одной стороне окрашен в *красный*

цвет, а на другой стороне — в *черный* цвет. Все n жетонов высыпаются из урны на поверхность стола. Обозначим через W_n^{red} (W_n^{black}) *сумму номеров на жетонах, выпавших красным (черным) цветом.* Тогда случайные величины W_n^{red} и W_n^{black} имеют одно и то же распределение вероятностей Вилкоксона, совпадающее с распределением W_n. Кроме того,

$$W_n^{red} + W_n^{black} = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

Пример 2. Пусть результаты пятнадцати испытуемых по некоторому тесту при двух последовательных замерах с равной вероятностью могут улучшиться или ухудшиться, причем показатель динамики (разность результатов тестирования) имеет одинаковое для всех испытуемых симметричное распределение. Заготовим для каждого испытуемого картонный жетон, две стороны которого раскрашенны в красный и черный цвет, надпишем сначала на обеих сторонах жетона показатель динамики данного испытуемого, а затем ранг абсолютной величины этого показателя в ряду испытуемых. Разложим жетоны, повернув каждый из них красной стороной, если тестовый результат данного испытуемого улучшился, и черной, если ухудшился.

Просуммируем ранги красных жетонов и получим *статистику* Вилкоксона для данной выборки.

Обратим внимание на то, что написанный на жетоне результат с равной вероятностью мог бы быть противоположным (в силу симметрии распределения результатов). Это показывает, что теоретическое распределение статистики Вилкоксона для рассматриваемой задачи совпадает со смоделированным в предыдущем примере распределением "жетонной" случайной величины, параметры которой легко могут быть вычислены с помощью теорем о математическом ожидании и дисперсии суммы независимых случайных величин.

Случайная величина W_n принимает целочисленные неотрицательные значения. Диапазон возможных значений, принимаемых этой величиной, а также математическое ожидание и дисперсия W_n задаются соотношениями

$$0 \le W_n \le \frac{n(n+1)}{2},$$

(0, если все жетоны черные, и $1 + \cdots + n = \frac{n(n+1)}{2}$, если все жетоны красные),

$$\mathbf{M}W_n = \frac{n(n+1)}{4}, \quad \mathbf{D}W_n = \frac{n(n+1)(2n+1)}{24}.$$

Кроме того, распределение Вилкоксона *симметрично* относительно среднего $\mathbf{M}W_n = \frac{n(n+1)}{4}$, т.е. для любого x, $0 \le x \le \frac{n(n+1)}{2}$ вероятность

$$P(W_n = x) = P\left(W_n = \frac{n(n+1)}{2} - x\right).$$

При $0 \le x \le \frac{n(n+1)}{4}$ вероятность $P(W_n = x)$ монотонно возрастает и достигает своего максимального значения при x, равном целой части числа $\frac{n(n+1)}{4}$. При $\frac{n(n+1)}{4} \le x \le \frac{n(n+1)}{2}$ вероятность $P(W_n = x)$ монотонно убывает.

Обозначим через w_α^-, w_α^+ — *нижний и верхний двусторонние квантили* распределения Вилкоксона, соответствующие уровню значимости α. Нижний квантиль w_α^- определяется неравенствами

$$P(W_n \le w_\alpha^-) \le \frac{\alpha}{2}, \quad P(W_n \le w_\alpha^- + 1) > \frac{\alpha}{2},$$

а аналогичные неравенства, определяющие верхний квантиль w_α^+, имеют вид

$$P(W_n \ge w_\alpha^+) \le \frac{\alpha}{2}, \quad P(W_n \ge w_\alpha^+ - 1) > \frac{\alpha}{2}.$$

В силу симметрии распределения Вилкоксона верхний и нижний квантили связаны равенством

$$w_\alpha^- + w_\alpha^+ = \frac{n(n+1)}{2}.$$

Упражнение 6.1. Убедиться, что приведенные формулы выражают ту же идею, что и правила нахождения квантилей биномиального распределения, приведенные в предыдущей главе в разделе, посвященном критерию знаков.

Численные значения нижних квантилей распределения Вилкоксона указаны в таблице В в конце книги.

6.1.1. Точечная оценка математического ожидания

Пусть $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ — исходные наблюдения выборки. Для каждой пары целых чисел (i, j), $1 \le i \le j \le n$, где число таких пар равно

$$N = \frac{n(n+1)}{2},$$

обозначим через

$$\Sigma_{ij} \stackrel{\text{def}}{=} \frac{x_i + x_j}{2}, \quad 1 \le i \le j \le n$$

полусумму соответствующих наблюдений. Эти полусуммы удобно представить в виде треугольной $n \times n$–таблицы (матрицы) $||\Sigma_{ij}||$, на диагонали которой выписаны элементы выборки $\Sigma_{ii} = x_i$, $i = 1, 2, \dots n$, а выше диагонали в соответствующих клетках располагаются полусуммы Σ_{ij}, $i < j$.

Запишем получившиеся полусуммы в возрастающем порядке

$$\Sigma^1 \le \Sigma^2 \le \dots \le \Sigma^N, \quad N = \frac{n(n+1)}{2},$$

т.е. образуем *вариационный ряд* из данных полусумм. Наиболее простой *графический способ* получения вариационного ряда состоит в *изображении* рассматриваемых полусумм на подходящим образом выбранной числовой оси.

В качестве *непараметрической точечной оценки* математического ожидания a рассматривается *медиана* $\hat{a}$ ряда полусумм, которая определяется как число наиболее близкое к центру данного вариационного ряда[1]. Более точно

$$\hat{a} = \text{med}\left\{\Sigma^1 \le \Sigma^2 \le \dots \le \Sigma^N\right\} \stackrel{\text{def}}{=} \begin{cases} \Sigma^{(N+1)/2}, & \text{если N нечётно,} \\ \frac{\Sigma^{N/2}+\Sigma^{N/2+1}}{2}, & \text{если N чётно,} \end{cases}$$

Преимущество медианы $\hat{a}$ по сравнению с выборочным средним $\bar{x}$ состоит в существенно меньшей зависимости медианы от *артефактов*, т.е. "слишком больших" (или "слишком малых") элементов выборки.

6.1.2. Непараметрический доверительный интервал математического ожидания

Если нижний двусторонний квантиль $w_\alpha^- \ge 1$ (*или верхний двусторонний квантиль* $w_\alpha^+ \le N-1$), *то* можно показать, что *для любого теоретического закона распределения выборки*

$$P(\Sigma^{w_\alpha^-+1} < a < \Sigma^{w_\alpha^+}) \ge 1 - \alpha.$$

Поэтому при коэффициенте доверия $1 - \alpha$ получаем следующий рецепт построения границ доверительного интервала для теоретического среднего значения a в непараметрической модели выборки:

$$a_\alpha^- = \Sigma^{w_\alpha^-+1}, \quad a_\alpha^+ = \Sigma^{w_\alpha^+}.$$

---

[1] Если количество членов вариационного ряда нечетно, берется средний член, если четно, то полусумма двух центральных членов.

Эти формулы означают, что левый конец доверительного интервала a_α^- является членом вариационного ряда полусумм с номером $w_\alpha^- + 1$, а правый конец доверительного интервала a_α^+ является членом вариационного ряда полусумм с номером w_α^+. Можно заметить, что правый конец доверительного интервала a_α^+ при обратном отсчёте с правого конца вариационного ряда имеет порядковый номер w_α^-.

Продолжение данного параграфа мы рекомендуем читать только тем, кто хочет лучше понять механизм построения доверительных интервалов.

Рассмотрим показатели экспериментальной группы нашего основного примера (который мы рассматривали в третьей и пятой главах) вместе с рангами их абсолютных величин:

−9	21	24	25	26	26	31	37	38	45
1	2	3	4	5	6	7	8	9	10

Статистика Вилкоксона в этом случае равна 1, что позволяет нам отвергнуть гипотезу о равенстве нулю соответствующего математического ожидания даже на уровне значимости 0,02 для двустороннего критерия, поскольку соответствующий квантиль равен 5.

Построение доверительного интервала эквивалентно ответу на аналогичный вопрос сразу для всех значений математического ожидания. В доверительный 0,02 —интервал входят те значения математического ожидания, которые не будут отвергнуты выбранным нами двусторонним[2] критерием Вилкоксона.

Каким же образом мы можем решить эту проблему сразу для всех значений?

Рассмотрим очень короткую выборку и отследим поведение ранговых сумм при разных вариантах среднего значения. Пусть троекратное испытание случайной величины X дало результат:

$$3, 7, 15$$

Если H_0 есть гипотеза $MX = 0$, то сумма рангов отрицательных членов выборки равна нулю. Если сформулировать гипотезу H_0': $MX = 4$,

[2] Доверительный интервал всегда строится по двусторонним квантилям. Получаемые доверительные границы согласуются только с соответствующим двусторонним критерием. Односторонний критерий на уровне значимости α может отвергнуть значение математического ожидания, которое попадает в доверительный интервал того же уровня значимости α.

то для ее проверки можно употребить тот же критерий, если воспользоваться им для оценки эквивалентной гипотезы $M(X-4)=0$.

Выборка, соответствующая случайной величине $X-4$, такова:

$$(3-4,\ 7-4,\ 15-4) \text{ или } (-1,\ 3,\ 11).$$

Для нее сумма отрицательных рангов равна уже единице. Как легко заметить, ранговая сумма превращается из нуля в единицу, если ставится вопрос о равенстве MX числу, немного превосходящему минимальный член исходной выборки, равный 3.

Следующее возрастание ранговой суммы произойдет после точки $5=(3+7)/2$: Например, для гипотезы $M(X-4{,}9)=0$ соответствующая выборка такова:

$$(3-4{,}9,\ 7-4{,}9,\ 15-4{,}9) \text{ или } (-1{,}9,\ 2{,}1,\ 10{,}1)$$

и ранговая сумма все еще единица, а для гипотезы $M(X-5{,}1)=0$ выборка такова:

$$(3-5{,}1,\ 7-5{,}1,\ 15-5{,}1) \text{ или } (-2{,}1,\ 1{,}9,\ 9{,}9).$$

Отрицательное число становится большим по модулю, чем ближайшее положительное, и сумма рангов принимает следующее значение, а именно 2.

Ранговая сумма будет увеличиваться далее в точках

$$(7+7)/2=7;\ (3+15)/2=9;\ (7+15)/2=11;\ (15+15)/2=15.$$

Всего, если размер выборки 3, будет наблюдаться 6 точек роста ранговой суммы для отрицательных значений, при переходе через которые она будет меняться от нуля до шести.

В общем случае для выборки размера n таких переходов будет $\frac{n(n+1)}{2}$. Если двусторонний нижний квантиль $w^{-}_{0{,}05}$ равен, предположим, 17, то будут отвергнуты гипотезы о равенстве нулю $M(X-a)$, пока при увеличении a (проще всего сказать: при увеличении a начиная от $-\infty$) не будет пройдено 17 точек возрастания ранговых сумм. После этого вопрос о гипотезе $M(X-a)=0$ будет решаться в ее пользу, пока a не превысит семнадцатую с конца точку роста, после чего ранговая сумма отрицательных значений станет аномально большой для того, чтобы могла быть принята гипотеза $M(X-a)=0$, и она будет отвергнута для всех больших значений a. Таким образом в доверительный пятипроцентный интервал будут включены все значения a,

лежащие между 17-й слева точкой возрастания ранговых сумм и 17-й справа такой точкой. Сами же точки возрастания, как мы убедились, представляют собой полусуммы пар не обязательно различных членов исходной выборки.

6.2. Распределение Манна—Уитни

Распределение Манна—Уитни формально задается следующим образом. Пусть $\gamma_1, \ldots \gamma_n$ и $\gamma'_1, \ldots \gamma'_m$ одинаково распределенные независимые случайные величины с нулевыми средними значениями. При этом их неизвестная функция распределения

$$F(t) \stackrel{\text{def}}{=} \Pr\{\gamma_i < t\} = \Pr\{\gamma'_j < t\}$$

предполагается *непрерывной*.

Рассмотрим случайные величины:

$$u_{\gamma\gamma'} = \text{число пар } (i,j), \text{ для которых } \gamma_i < \gamma'_j \quad (\gamma'_j - \gamma_i > 0),$$

$$u_{\gamma'\gamma} = \text{число пар } (i,j), \text{ для которых } \gamma_i > \gamma'_j \quad (\gamma'_j - \gamma_i < 0).$$

Замечание 1. Случай равенства элементов в паре для непрерывной случайной величины имеет нулевую вероятность. На практике критерий используется и при равенстве значений в сравниваемых выборках. По правилу, вполне аналогичному введенному ранее для знаково-рангового критерия, пара с равными значениями дает вклад $1/2$ в каждую сумму $u_{\gamma\gamma'}$ и $u_{\gamma'\gamma}$

Поскольку общее число возможных пар, содержащих по одному элементу каждой выборки, равно mn, то очевидно $u_{\gamma\gamma'} + u_{\gamma'\gamma} = nm$.

Можно показать, что распределение вероятностей случайной величины $u_{\gamma\gamma'}$ совпадает с распределением вероятностей $u_{\gamma'\gamma}$ и их общее распределение не зависит от распределения вероятностей случайных величин γ и γ', а зависит лишь от объёмов выборок n и m. Мы будем обозначать случайную величину, имеющую распределение Манна—Уитни, U_n^m.

Пример 3. В урне имеются n красных и m черных жетонов. Все $n+m$ жетонов выкладываются из урны случайным образом слева направо в последовательность длины $n+m$. Пусть $1 \leq r_1 < r_2 < \cdots < r_n$, где $i \leq r_i \leq n+m$, $i = 1, 2, \ldots, n$, обозначают записанные в возрастающем

порядке номера позиций (ранги) красных жетонов, а $1 \le b_1 < b_2 < \dots < b_m$, где $j \le b_j \le n + m$, $j = 1, 2, \dots, m$ обозначают записанные в возрастающем порядке номера позиций (ранги) черных жетонов в данной последовательности. Отметим, что всегда $r_i \neq b_j$.

Рассмотрим случайную величину: "количество всевозможных пар жетонов, в которых слева красный жетон, а справа черный", или

$$u_{rb} \stackrel{\text{def}}{=} \text{число пар } (i, j), \text{ для которых } r_i < b_j.$$

Эта случайная величина является модельным примером для распределения Манна—Уитни U_n^m.

В силу симметрии такое же распределение имеет случайная величина "количество всевозможных пар жетонов, в которых слева черный жетон, а справа красный", или

$$u_{br} \stackrel{\text{def}}{=} \text{число пар } (i, j), \text{ для которых } r_i > b_j, = nm - u_{rb}$$

То же самое количество пар можно считать по другой формуле:

$$u_{rb} = \sum_{j=1}^{m} b_j - \frac{m(m+1)}{2}.$$

Действительно, ранг любого черного жетона на единицу больше, чем количество жетонов, расположенных слева от него. На языке пар жетонов это количество представляет собой сумму двух слагаемых — количества пар, в которых слева красный, а справа данный черный жетон, и количества пар, где слева от данного черного жетона помещен черный. Это последнее количество равно $C_m^2 = \frac{m(m-1)}{2}$. Чтобы получить сумму рангов черных жетонов надо еще прибавить по единице на каждый жетон (см. начало абзаца). Поскольку $\frac{m(m-1)}{2} + m = \frac{m(m+1)}{2}$, то окончательно

$$\sum_{j=1}^{m} b_j = u_{rb} + \frac{m(m+1)}{2},$$

откуда

$$u_{rb} = \sum_{j=1}^{m} b_j - \frac{m(m+1)}{2}.$$

Разумеется, верна и симметричная формула

$$u_{br} = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2}.$$

Пример 4. Пусть $n = 5$, $m = 4$, а получившаяся случайная последовательность длины $n + m = 9$ имеет вид: $r\ r\ b\ r\ b\ r\ b\ b\ r$, где символами "$r$" обозначены красные жетоны, а символами "b" — чёрные. Тогда ранги красных жетонов: $r_1 = 1$, $r_2 = 2$, $r_3 = 4$, $r_4 = 6$, $r_5 = 9$, а ранги черных жетонов: $b_1 = 3$, $b_2 = 5$, $b_3 = 7$, $b_4 = 8$. При этом $u_{rb} = 13$, $u_{br} = 7$.

Диапазон значений случайной величины U_n^m, математическое ожидание и дисперсия записываются в виде:

$$0 \le U_n^m \le nm, \qquad MU_n^m = \frac{nm}{2}, \qquad DU_n^m = \frac{nm(n+m+1)}{12}.$$

Вероятности из распределения Манна—Уитни обладают свойством симметрии относительно математического ожидания $MU_n^m = \frac{nm}{2}$, т.е. вероятность

$$P(U_n^m = x) = P\left(U_n^m = nm - x\right) \quad \text{для любого} \quad x = 0, 1, 2, \ldots, nm.$$

Кроме того, вероятность $P(U_n^m = x)$ монотонно возрастает при $0 \le x \le nm/2$ и монотонно убывает при $nm/2 \le x \le nm$.

$P(U_n^m = x)$ достигает наибольшего значения при x, ближайшем к $nm/2$.

6.2.1. Квантили распределения Манна—Уитни

Для уровня значимости α *нижний односторонний квантиль* u_α^- распределения Манна—Уитни задается условиями:

$$P(U_n^m \le u_\alpha^-) \le \alpha, \quad P(U_n^m \le u_\alpha^- + 1) > \alpha,$$

т.е., двигаясь слева направо (от $-\infty$), мы выбираем последнюю точку, для которой вероятность того, что имеющая распределение U_n^m случайная величина попадет в нее или левее, все еще меньше или равна α.

Аналогично *верхний односторонний квантиль* u_α^+ задается условиями

$$P(U_n^m \ge u_\alpha^+) \le \alpha, \quad P(U_n^m \ge u_\alpha^+ - 1) > \alpha,$$

т.е. это первая при движении слева направо точка, для которой вероятность попадания в нее или правее становится меньше либо равной α.

Поскольку $P(u_\alpha^- < U_n^m < u_\alpha^+) \ge 1 - 2\alpha$, то введенные квантили являются также *двусторонними квантилями для уровня значимости* 2α.

В силу симметрии распределения Манна—Уитни относительно математического ожидания $\frac{nm}{2}$

$$u_\alpha^+ - \frac{nm}{2} = \frac{nm}{2} - u_\alpha^-, \text{ поэтому } u_\alpha^- + u_\alpha^+ = nm.$$

6.2.2. Точечная оценка теоретического сдвига $\theta = b - a$

Пусть случайные величины X и Y представляют собой сдвиги одной и той же случайной величины γ, выраженные формулами $X = a + \gamma$ и $Y = b + \gamma$, а γ, как и прежде, имеет нулевое математическое ожидание.

Пусть $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ и $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ — наблюдения, полученные независимыми испытаниями соответственно X и Y и требуется оценить неизвестную разность математических ожиданий. Мы уже умеем оценивать разность математических ожиданий (или, что одно и то же, математическое ожидание разности) $M_Y - M_X$ разностью средних арифметических $\bar{y} - \bar{x}$. Возможен другой способ.

Рассмотрим разности Δ_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, число которых равно nm. Их удобно представить в виде прямоугольной $(n \times m)$ — таблицы (матрицы) $\|\Delta_{ij}\|$. Отметим, что значение статистики Манна—Уитни u_{xy} (u_{yx}) есть количество положительных (отрицательных) элементов матрицы $\|\Delta_{ij}\|$ плюс число, равное половине количества нулевых элементов матрицы $\|\Delta_{ij}\|$.

Образуем из разностей Δ_{ij} *вариационный ряд*, записывая их в возрастающем порядке:

$$\Delta^1 \leq \Delta^2 \leq \Delta^3 \leq \cdots \leq \Delta^{nm}.$$

В качестве непараметрической точечной оценки параметра $\theta = b - a$ рассматривается *медиана* получившегося вариационного ряда:

$$\hat{\theta} = \mathrm{med}\,\{\Delta^1 \leq \Delta^2 \leq \Delta^3 \leq \cdots \leq \Delta^{nm}\} \stackrel{\text{def}}{=} \begin{cases} \Delta^{(nm+1)/2}, & \text{если } nm \text{ — нечетно} \\ \frac{\Delta^{nm/2} + \Delta^{nm/2+1}}{2}, & \text{если } nm \text{ — четно.} \end{cases}$$

Оценка сдвига медианой имеет преимущество перед оценкой с помощью средних арифметических в том случае, если выборки могут включать аномальные результаты. В реальной работе такие случаи весьма часты. В жизненных ситуациях очень трудно добиться однородности экспериментальных групп, и часто приходится использовать "то, что есть". В подобных случаях, если сравниваются, например, результаты двух типов какого-либо обучающего воздействия в двух естественных

группах (например школьных классах), то наличие в одной из них гениального ученика приводит к искажению результатов. Метод оценки с помощью медианы более устойчив к такому искажению, поскольку очень большие или очень маленькие значения не приводят к существенному смещению оценок, как это происходит со средним арифметическим.

6.2.3. Доверительный интервал для сдвига средних

Пусть при коэффициенте доверия $1-\alpha$ нижний двусторонний квантиль $u_\alpha^- \geq 1$ или верхний двусторонний квантиль $u_\alpha^+ \leq nm - 1$. Тогда можно показать, что для любой функции распределения вероятностей ошибки $F(t)$ справедливо неравенство

$$P(\Delta^{u_\alpha^- + 1} < \theta < \Delta^{u_\alpha^+}) \geq 1 - \alpha.$$

Поэтому при коэффициенте доверия $1 - \alpha$ имеем следующий рецепт построения границ доверительного интервала для сдвига в непараметрической модели

$$\theta_\alpha^- = \Delta^{u_\alpha^- + 1}, \quad \theta_\alpha^+ = \Delta^{u_\alpha^+}.$$

Смысл данных оценок доверительных границ совершенно аналогичен описанному в разделе о непараметрическом доверительном интервале для знаково-рангового критерия, и продвинутый читатель может самостоятельно справиться с задачей его прояснения.

Глава 7

Гипотезы о связи случайных величин

Задача о связи характеристик изучаемого предмета — одна из наиболее часто встречающихся в психологических исследованиях. Первый, самый грубый, но и, вероятно, самый надежный ответ на вопрос о связи дает расчет корреляции и родственных показателей, которые мы рассмотрим в начале данной главы. Затем мы коснемся другого способа описания связи характеристик — линейной регрессии.

7.1. Корреляция случайных величин. Коэффициент Фишера—Пирсона

В разделе 7.2 первой части книги корреляция уже упоминалась в контексте факторного анализа. Там же были приведены примеры диаграмм рассеяния выборок, состоящих из пар наблюдений, соответствующих разным значениям выборочной корреляции (рис. 7.1, 7.2, 7.3, 7.4).

Типичная корреляционная задача возникает в исследованиях, где каждый испытуемый характеризуется двумя или большим числом показателей. Пусть, например, n испытуемых студентов пишут две контрольные работы по общей психологии и математике. Результат i-го студента записывается в виде пары чисел (x_i, y_i). Ставится вопрос, имеется ли связь между результатами контрольных по математике и психологии?

Вычислим сначала выборочные оценки $\bar{x}$, S_x^2, $\bar{y}$ и S_y^2, по формулам

$$\bar{x} = (x_1 + \cdots + x_n)/n,$$

$$S_x^2 = ((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)/n$$

(здесь следует брать смещенную оценку дисперсии). Аналогично для $\bar{y}$.

Перейдем к *стандартизованным* выборкам. Для этого каждый показатель x_i преобразуем по формуле

$$x_i' = \frac{x_i - \bar{x}}{s_x}$$

(аналогично для $\bar{y}$). Если теперь подсчитать оценки среднего и среднеквадратического отклонения для стандартизованных переменных, то в силу алгебраических тождеств окажется $\bar{x}' = \bar{y}' = 0$, $s_x = s_y = 1$.

Если имеет место связь между исходными переменными, то и стандартизованные переменные будут ее демонстрировать.

Опустим в наших обозначениях штрихи и будем считать, что переменные x_i и y_i стандартизованы.

Показатели корреляции улавливают *линейную* зависимость между переменными, например, такую:

— "чем больше x, тем, скорее всего, больше y", или

— "чем больше x, тем, скорее всего, меньше y".

Если имеет место первая из них, то положительным значениям x (в нашем примере более высоким, чем средние, оценкам по математике) будут соответствовать положительные значения y (более высокие, чем средние, оценки по психологии). И, наоборот, отрицательным значениям x, скорее всего, будут соответствовать отрицательные значения y того же испытуемого. В этом случае произведение $x_i y_i$, скорее всего, окажется положительным для всякого испытуемого.

Если имеет место вторая, "негативная" зависимость, то произведения $x_i y_i$, скорее всего, окажутся отрицательными.

Составим сумму

$$r_{xy} = \frac{1}{n}(x_1 y_1 + \cdots + x_n y_n).$$

И будем рассматривать ее как меру связи между данными характеристиками у наших испытуемых. В силу алгебраических причин[1] для

[1] Для знакомых с линейной алгеброй эти причины вполне прозрачны: r_{xy} это скалярное произведение векторов.

стандартизованных выборок сумма r_{xy} всегда заключена между -1 и $+1$, причем -1 означает строгую линейную зависимость $y_i = -x_i$, а 1 — зависимость $y_i = x_i$. Промежуточные значения r_{xy} соответствуют более или менее выраженной зависимости, на которую накладываются случайные вариации переменных. Значению $r_{xy} = 0$ соответствует отсутствие зависимости между результатами испытаний.

Показатель r_{xy} называется *выборочным коэффициентом корреляции Фишера* (в литературе и компьютерных пакетах этот коэффициент называют также коэффициентом корреляции Пирсона).

Если пары испытаний физически независимы, то выборочный коэффициент корреляции тем не менее не будет стабильным нулем, а будет колебаться вокруг нуля.

Упражнение 7.1. Возьмите две симметричных монеты достоинством в одну копейку и один европейский цент. Проведите серию из пяти подбрасываний пары монет, и запишите результаты в виде $(x_1, y_1), \ldots, (x_5, y_5)$, полагая

$x_i = 1$, если копейка выпала стороной "цифра",

$x_i = 0$, если копейка выпала "гербом",

$y_i = 1$, если цент выпал стороной "цифра",

$y_i = 0$, если цент выпал "гербом".

Проведите далее стандартизацию выборок и подсчитайте коэффициент корреляции.

Повторите процедуру несколько раз и убедитесь, что нулевое значение выборочного коэффициента корреляции явление весьма редкое. При многократном повторении опыта можно убедиться, что его результат имеет некоторое *распределение*. В следующем разделе мы будем говорить о распределении выборочного коэффициента корреляции именно в этом смысле.

7.1.1. Проверка гипотезы о корреляционной зависимости

Если X и Y независимые нормально распределенные случайные величины, $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ — выборки, представляющие собой результаты независимых испытаний этих случайных величин, а $x'_1, \ldots, x'_n$ и $y'_1, \ldots, y'_n$ соответствующие стандартизованные выборки, то распределение выборочного коэффициента корреляции, подсчитанного по стандартизованным выборкам, не зависит от параметров нормального распределения случайных величин X и Y.

Это значит, что квантили распределения выборочного коэффициента корреляции Фишера зависят только от уровня значимости и размера выборки n. В верхней строке нижеследующей таблицы приведены квантили распределения выборочного коэффициента корреляции по Фишеру и Спирмену для $n = 10$.

Модель	α	0,05	0,025	0,01	0,005
Фишер	$r_\alpha(10)$	0,497	0,576	0,658	0,709
Спирмен	$\hat{r}_\alpha(10)$	0,564	0,648	0,745	0,794

Можно заметить, что если верна гипотеза H_0 об отсутствии зависимости между случайными величинами, то выборочный коэффициент при $n = 10$ может принимать тем не менее довольно большие значения, так что даже пятипроцентный квантиль требует для принятия гипотезы о зависимости случайных величин, чтобы выборочный коэффициент достигал почти 0,5.

7.2. Корреляция случайных величин. Коэффициент Спирмена

Если предположение о нормальности случайных величин X и Y, в результате испытаний которых были получены парные выборки $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ не соответствует действительности, то для проверки гипотез о связи следует применять непараметрические методы. Наиболее употребительный из них — коэффициент корреляции Спирмена. Для его расчета запишем наблюдения $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ в порядке возрастания, т.е. образуем из результатов измерений $x_1, x_2, \ldots, x_n$ *вариационный ряд* и поставим в соответствие измерению x_i его номер (ранг) в этом ряду $R_i(\mathrm{x})$.

Если число x_i встречается среди наблюдений x два или более раз, то его рангом называется *среднее арифметическое значение порядковых номеров* членов вариационного ряда, которые совпадают с x_i.

Очевидно, что при таком определении ранга $R_i(\mathrm{x})$ сумма всех рангов

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Аналогичным образом определяются ранги $R_i(\mathrm{y})$ для наблюдений $\mathrm{y} = (y_1, y_2, \ldots, y_n)$.

Введём вычисляемую по n парам (x_i, y_i) статистику

$$\hat{r}_{xy} \stackrel{\text{def}}{=} 1 - \frac{6S}{n^3 - n}, \quad \text{где} \quad S \stackrel{\text{def}}{=} \sum_{i=1}^{n} (R_i(\mathrm{x}) - R_i(\mathrm{y}))^2,$$

называемую *коэффициентом ранговой корреляции Спирмена.*

Можно показать, что при любых значениях пар (x_i, y_i) число $\hat{r}_{xy}$ лежит в отрезке $[-1; +1]$, т.е. $-1 \le \hat{r}_n \le 1$.

Упражнение 7.2. Проверить, что для выборок (x: 1, 2, 3), (y: 1, 2, 3) выборочный коэффициент корреляции Спирмена равен 1, а для выборок (x: 1, 2, 3), (y: 3, 2, 1) и (x: 1, 2, 3, 4), (y: 4, 3, 2, 1) коэффициент $\hat{r}_{xy}$ равен -1.

Для проверки гипотезы H_0 о независимости случайных величин X и Y для непараметрической модели Спирмена полученный выборочный коэффициент надо сравнить с соответствующими квантилями. Для $n = 10$ квантили можно найти в таблице предыдущего раздела. Можно заметить, что спирменовский коэффициент при независимости выборок подвержен еще большим колебаниям, чем коэффициент Фишера, и для принятия гипотезы о наличии связи требуется относительно большее значение $\hat{r}_{xy}$, чем r_{xy}.

Весьма полезно будет следующее

Упражнение 7.3. Измерьте рост и вес 10 своих товарищей, составьте соответствующие выборки и вычислите коэффициенты r_{xy} и $\hat{r}_{xy}$. Проверьте гипотезу о независимости веса и роста человека по вашим выборкам по таблице предыдущего раздела. На каких уровнях значимости вы можете отвергнуть гипотезу о независимости роста и веса?

7.3. Корреляция случайных величин. Таблицы сопряженности

Метод Спирмена перестает надежно работать, если в выборках имеется большое количество повторяющихся значений. В этом случае остается использовать самый универсальный метод проверки гипотез о связи признаков — метод таблиц сопряженности.

Мы рассмотрим здесь самый простой вариант их использования — так называемые таблицы 2×2.

Пример 1. Рассмотрим случайный опыт, описанный в упражнении 7.1 (подбрасывание копейки и цента), предположив дополнительно,

что монеты изогнуты и вероятности “герба” и “цифры” для копейки равны p_k и q_k, а для цента — p_c и q_c. В силу независимости результатов падения монет вероятности совместного наступления событий можно задать таблицей

	U = цент: “герб”	$\bar{U}$ = цент: “цифра”
V = копейка: “герб”	$p(UV) = p_c p_k$	$p(\bar{U}V) = q_c p_k$
$\bar{V}$ = копейка: “цифра”	$p(U\bar{V}) = p_c q_k$	$p(\bar{U}\bar{V}) = q_c q_k$

Обратим внимание на то, что $p(UV)/p(U\bar{V}) = p(\bar{U}V)/p(\bar{U}\bar{V}) = p_k/q_k$, откуда следует, что

$$p(UV)p(\bar{U}\bar{V}) = p(U\bar{V})p(\bar{U}V).$$

Это частный случай весьма важного общего факта: таблица, описывающая вероятности сочетаний двух независимых событий, каждое из которых имеет два исхода, всегда обладает свойством: произведения элементов по диагоналям равны.

Если теперь мы проведем большую серию опытов и запишем в аналогичную таблицу частоты наступления пар событий, то, поскольку эти частоты приближаются к соответствующим вероятностям, равенство диагональных произведений будет приблизительно соблюдаться и для наблюдаемых частот. Если же пары событий не являются независимыми, то равенство будет заметно нарушаться.

Применим это свойство к задаче о связи признаков.

Пусть, как и прежде, мы имеем набор пар наблюдений (x_i, y_i), причем среди них возможно значительное число повторяющихся. Найдем такое граничное значение m_x, чтобы выборка x_i разбивалась им примерно на равные части, нижнюю, содержащую $x_i < m_x$, и верхнюю, в которую входят $x_i > m_x$. Равенство не допускается, поэтому разбиение надо осуществлять с помощью числа m_x, не входящего в выборку.

Аналогично разобьем выборку y_i с помощью границы m_y.

Для каждой пары (x_i, y_i) имеется ровно четыре возможности:

$(A) : x_i < m_x,\ y_i < m_y$;

$(B) : x_i > m_x,\ y_i < m_y$;

$(C) : x_i < m_x,\ y_i > m_y$;

$(D) : x_i > m_x,\ y_i > m_y$.

Разобьем выборку на четыре соответствующие группы и, подсчитав количество членов выборки, попавших в каждую группу, обозначим их теми же буквами A, B, C, D.

Составим таблицу 2×2

A	B
C	D

Если верна гипотеза о независимости признаков H_0, то AD будет приблизительно равно CB, если же имеет место связь признаков, то разность $AD - CB$ будет заметно отличаться от нуля.

Для того чтобы решить вопрос о "заметности" или "незаметности" такого отклонения, необходимо задать статистику.

Имеется несколько возможностей, из которых упомянем две:

1) Статистика

$$\chi^2 = \frac{n(AD - CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

при достаточно больших n имеет распределение χ^2 с одной степенью свободы, и для принятия решения надо воспользоваться соответствующими квантилями: $\chi^+_{0,05} = 3{,}84$, $\chi^+_{0,01} = 6{,}635$ и $\chi^+_{0,001} = 10{,}83$. (Заметим, что в числителе стоит квадрат той самой разности, которая должна быть близка к нулю при независимости признаков, поэтому гипотеза H_0 отвергается, если квадрат этой разности аномально велик, т.е. статистика превосходит соответствующий квантиль.)

2) Воспользоваться непосредственно статистическими таблицами для 2×2 таблиц сопряженности, которые дают наиболее точный результат. В силу громоздкости этих таблиц, мы их приводить здесь не будем.

7.4. Линейный регрессионный анализ

7.4.1. Определение регрессионной прямой

Предположим, социологический опрос n мужчин в возрасте от 20 до 40 лет включал два вопроса: (1) возраст, (2) порядковый номер последнего брака, в котором состоял (состоит) опрашиваемый.

Полученные результаты отображают n пар чисел (t_i, x_i), где t_i — возраст опрошенного, а x_i — номер последнего брака.

Понятно, что показатель количества браков в среднем зависит от возраста. Линейный регрессионный анализ исходит из гипотезы, что эта зависимость линейна и показатель количества браков представляет

собой сумму детерминированного вклада линейной функции от возраста[2] и случайного вклада, зависящего от индивидуальных особенностей данного опрошенного мужчины.

Таким образом, предполагается зависимость $x_i = c_1 t_i + c_0 + \gamma_i$, где c_1 и c_0 некоторые коэффициенты, а γ_i случайная добавка, для каждого испытуемого своя.

Ясно, что задача не имеет очевидного однозначного решения. Всякая прямая может фигурировать в правой части равенства, от ее выбора будет зависеть только компенсирующая "добавка" γ. Если считать, что γ_i суть независимые испытания некоторой случайной величины, характеризующей какие-то специфические качества испытуемых, то можно попытаться найти такую прямую, чтобы выборочная дисперсия этой величины оказалась минимальной из всех возможных. Это означает минимизацию суммы $\gamma_1^2 + \gamma_2^2 + \cdots + \gamma_n^2$.

Операция по нахождению наилучшей в каком-то смысле прямой, описывающей подобную зависимость, называется *линейным регрессионным анализом*, а метод, обеспечивающий одно из возможных решений, — то, которое характеризуется минимумом дисперсии случайной составляющей, называется методом *наименьших квадратов Гаусса*[3]. Мы дадим строгие выкладки в следующей главе, а здесь перейдем непосредственно к результатам.

Вычисления значительно упрощаются, если сделать полезную замену переменной и измерять возраст не в годах от рождения, а в годах относительно среднего возраста испытуемых, т.е. $t'_i = t_i - \bar{t}$.

Вычислим выборочную ковариацию, которая после замены задается простой формулой

$$R_{tx} = \frac{1}{n}(t_1 x_1 + \cdots + t_n x_n),$$

выборочное среднее по x

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

[2] Естественно, что такая модель работает только в ограниченных пределах: для очень молодых и очень старых мужчин количество браков растет с заведомо меньшей скоростью, чем у лиц, попавших в данную выборку. Тем не менее для предсказания значения показателя в известных возрастных пределах такой метод вполне оправдан.

[3] Этот метод применяется не только к задачам регрессионного анализа. Другой пример его употребления рассматривался в примере 3 седьмой главы первой части книги.

и (смещенную) оценку дисперсии выборки $(t'_1, \ldots, t'_n)$

$$S_t^2 = \frac{1}{n}((t'_1)^2 + \cdots + (t'_n)^2).$$

Параметры регрессионной прямой можно найти после этого по формулам

$$\hat{c}_1 = \frac{R_{tx}}{S_t^2}, \quad \hat{c}_0 = \bar{x},$$

т.е. наилучшей в смысле минимума суммы квадратов отклонений является прямая

$$x = \frac{R_{tx}}{S_t^2} t + \bar{x},$$

а зависимая переменная (в нашем примере порядковый номер последнего брака) выражается через модифицированный возраст по формуле

$$x_i = \frac{R_{tx}}{S_t^2} t_i + \bar{x} + \gamma_i.$$

Замечание 1. Как и в примере 3 первой главы второй части книги, мы нашли, что результат использования метода наименьших квадратов указывает на среднее арифметическое переменных. Наша регрессионная прямая проходит через точку $(\bar{t}, \bar{x})$.

Замечание 2. Более употребительные формулы корреляции и ковариации приведены в начале следующей главы.

Глава 8

Гипотезы о связи случайных величин (окончание)

8.1. Корреляция между случайными величинами

Пусть X и Y дискретные случаные величины, распределение которых задано таблицами

X	-1	1
	0,5	0,5

Y	-1	1
	0,5	0,5

Если X и Y независимы, то вероятность одновременного наступления событий $X = -1$, $Y = -1$ равна произведению вероятностей наступления каждого из событий в отдельности, т.е. $1/4$ (аналогично для остальных пар событий ($X = -1$, $Y = 1$), ($X = 1$, $Y = -1$) и ($X = 1$, $Y = 1$)). Однако так бывает далеко не всегда. Например, если $X = 1$ кодирует событие "вес испытуемого выше среднего", а $X = -1$ — "вес ниже среднего", а случайная величина Y кодирует аналогичные высказывания про рост испытуемого, то понятно, что вероятность одновременного превышения среднего роста и среднего веса выше, чем вероятность наблюдать большой вес при маленьком росте. В этом случае

имеет смысл говорить о совместном распределении случайных величин. Предположим, что для нашего примера оно может быть задано таблицей

$X=-1, Y=-1$	$X=-1, Y=1$	$X=1, Y=-1$	$X=1, Y=1$
0,4	0,1	0,1	0,4

Характеристикой степени отклонения этого распределения от распределения независимых величин служит *корреляция.*

Сначала вычислим *ковариацию* R_{XY} по следующей формуле:

$$\begin{gathered}R_{XY} = \\ = ((-1) - MX)((-1) - MY)\,p(X=-1, Y=-1) + \\ +((-1) - MX)((+1) - MY)\,p(X=-1, Y=1) + \\ +((+1) - MX)((-1) - MY)\,p(X=1, Y=-1) + \\ +((+1) - MX)((+1) - MY)\,p(X=1, Y=1)\,.\end{gathered}$$

Подставляя значения вероятностей и учитывая, что $MX = MY = 0$, получаем

$$R_{XY} = (-1)(-1)\cdot 0{,}4 + (-1)(1)\cdot 0{,}1 + (1)(-1)\cdot 0{,}1 + (1)(1)\cdot 0{,}4 = 0{,}6\,.$$

Упражнение 8.1. Подставить в предыдущую формулу вероятности 0,25, соответствующие независимости случайных величин, и убедиться, что значение ковариации в этом случае равно нулю.

Для того чтобы получить корреляцию ρ_{XY}, надо поделить ковариацию на среднеквадратические отклонения

$$\rho_{XY} = \frac{R_{XY}}{\sqrt{DX}\sqrt{DY}}.$$

В нашем случае $DX = DY = 1$. Подставляя в формулу получаем

$$\rho_{XY} = 0{,}6.$$

В общем случае для случайных величин, заданных таблицами

X	x_1	x_2	$\ldots$	x_m
	p_1	p_2	$\ldots$	p_m

Y	y_1	y_2	$\dots$	y_m
	q_1	q_2	$\dots$	q_m

,

ковариация и корреляция задаются формулами

$$R_{XY} = \sum_{i,j} (x_i - Mx)(y_j - My)\, p(X = x_i \text{ и } Y = y_j),$$

$$\rho_{XY} = \frac{R_{XY}}{\sigma_X \sigma_Y},$$

где $\sigma_X = \sqrt{DX}$, $\sigma_Y = \sqrt{DY}$.

Замечание 1. Если вместо Y взять саму X, то формула ковариации превратится в формулу дисперсии X. Действительно, если $i \neq j$, то $p(X = x_i \text{ и } X = x_j) = 0$, поскольку случайная величина не может одновременно принимать два разных значения; если $i = j$, то $p(X = x_i \text{ и } X = x_i)$ тавтологично равно $p(X = x_i)$.

Таким образом, в формуле для ковариации будут отличны от нуля только "диагональные" члены, для которых $i = j$, т.е.

$$R_{XX} = \sum_{i=j} (x_i - MX)(x_j - MX) p(X = x_i) = \sum_i (x_i - MX)^2 p_i = DX.$$

Следовательно, "похожесть" двух случайных величин характеризуется "похожестью" ковариации на их дисперсии.

Аналогичные формулы для непрерывных случайных величин мы не будем здесь приводить.

8.2. Преобразование Фишера

Пусть по-прежнему $(x_1, \dots, x_n)$ и $(y_1, \dots, y_n)$ выборки, характеризующие значение некоторых признаков у группы из n испытуемых.

Первой характеристикой связи признаков служит выборочная ковариация, которая вычисляется по формуле

$$R_{xy} = \frac{1}{n}((x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})).$$

Выборочная ковариация входит в ряд важных формул, но для наших целей оценки степени связи признаков обладает существенным недостатком — она зависит от масштаба переменных.

Упражнение 8.2. Пусть рост трех испытуемых выражается числами 180, 190, 200 сантиметров, а их вес соответственно 50, 60, 70 килограммов. Вычислить ковариацию и показать, что она изменится, если рост измерять в метрах.

Для того чтобы получить безразмерную величину, надо ковариацию поделить на выборочные средние квадратические отклонения переменных

$$s_x = \sqrt{(\tfrac{1}{n}((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2},$$

$$s_y = \sqrt{(\tfrac{1}{n}((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}.$$

Получаем формулу выборочной корреляции в наиболее часто употребляемом виде:

$$r_{xy} = \frac{R_{xy}}{s_x s_y}.$$

Если размер выборок достаточно велик ($n > 20$), то, кроме разобранного в предыдущей главе, возможен еще один способ оценивания значимости выборочной корреляции, в некоторых отношениях более удобный.

Если корреляция реальных случайных величин, в результате испытаний которых были получены выборки $(x_1, \ldots, x_n)$ и $(y_1, \ldots, y_n)$, равна некоторому числу ρ, то распределение статистики (напомним, что статистика является случайной величиной)

$$z_{xy} = \frac{1}{2} \ln \frac{1 + r_{xy}}{1 - r_{xy}}$$

имеет нормальное распределение с математическим ожиданием

$$Mz = \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} + \frac{\rho}{2n - 2},$$

и дисперсией

$$Dz = \frac{1}{n - 3}.$$

Переход от r_{xy} к z_{xy} называется преобразованием Фишера. Это преобразование дает возможность решать новые задачи.

В предыдущей главе был описан способ проверки гипотезы о равенстве нулю реальной корреляции случайных величин, в результате испытаний которых были получены выборки $(x_1, \ldots, x_n)$ и $(y_1, \ldots, y_n)$.

Он возможен, поскольку затабулировано распределение выборочной корреляции при условии, что реальная корреляция ρ равна нулю.

Если мы захотим по нашим выборкам проверить гипотезу о том, что реальная корреляция ρ равна, скажем, 0,2, то нам необходимо затабулировать распределения выборочных корреляций r_{xy} для $\rho = 0{,}2$ и так же для всех возможных значений ρ от -1 до 1, которые могут нас заинтересовать.

Трудность состоит в том, что эти распределения при разных значениях ρ совершенно различны, поэтому в каждом случае работа должна производиться заново, что делает ее совершенно необозримой. Поэтому прямой способ используется только для проверки гипотезы $H_0 : \rho = 0$ против односторонней или двусторонней альтернативы.

Преобразование Фишера приводит к тому, что при любом ρ мы можем использовать одни и те же квантили нормального распределения.

Особенно важное применение преобразования Фишера — сравнение двух выборочных коэффициентов корреляции. (Например, мы хотим исследовать вопрос, одинаковы ли корреляции между ростом и весом у жителей Москвы и Лиссабона.)

Пусть r_1 и r_2 выборочные коэффициенты корреляции, полученные по выборкам размера n_1 и n_2 соответственно.

Если реальные корреляции ρ_1 и ρ_2 равны, то случайная величина $z_1 - z_2$ будет иметь нормальное распределение с нулевым математическим ожиданием и дисперсией, равной сумме дисперсий Dz_1 и Dz_2, т.е.

$$\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}.$$

Тогда

$$z_{1-2} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

будет стандартной нормальной случайной величиной, и гипотеза о равенстве корреляций отвергается с помощью квантилей нормального распределения, если z_{1-2} принимает аномально большие (положительные или отрицательные) значения.

Упражнение 8.3. Пренебрегая в формуле

$$Mz = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2n-2}$$

последним слагаемым (которое при больших n невелико) составить алгоритм нахождения по данной выборке доверительного интервала для значения реальной корреляции.

8.3. Линейный регрессионный анализ. Построение регрессионной прямой методом Гаусса

Пусть в результате эксперимента получены n пар чисел

$$(t_1, x_1), \quad (t_2, x_2), \quad \ldots \quad (t_n, x_n).$$

Каждую пару чисел из (t_i, x_i) можно рассматривать как точку на плоскости в системе координат (t, x).

В методе наименьших квадратов (МНК) вычисляется регрессионная прямая, для которой сумма квадратов расстояний *по оси ординат* от точек (t_i, x_i) до регрессионной прямой минимальна. Угловой наклон $\hat{c}_1$ и сдвиг $\hat{c}_0$, называемые оценками МНК, определяются как значения параметров c_1 и c_0, для которых достигается

$$\min_{(c_0, c_1)} \left\{ \sum_{i=1}^{n} (x_i - c_0 - c_1 t_i)^2 \right\}.$$

Сделаем замену переменной $t' = t - \bar{t}$. Ясно, что наилучшая прямая зависит только от конфигурации точек на плоскости, хотя ее уравнение при замене переменных изменяется. Однако после получения коэффициентов в новых координатах мы легко сможем вернуться к старым.

Заметим прежде всего, что

$$\bar{t'} = \sum t'_i = \sum (t_i - \bar{t}) = \sum t_i - n\bar{t} = n\bar{t} - n\bar{t} = 0$$

(поскольку все суммирования производятся в пределах от 1 до n, обозначения пределов суммирования для краткости опускаем, опускаем также далее штрихи при переменной t, отметив только, что теперь $\bar{t} = 0$ и $\sum t_i = 0$).

В новых координатах формула ковариации значительно упрощается:

$$R_{tx} = \frac{1}{n} \sum (x_i - \bar{x})(t_i - \bar{t}) = \frac{1}{n} \sum (x_i - \bar{x}) t_i = \frac{1}{n} \sum x_i t_i - \frac{1}{n} \bar{x} \sum t_i = \frac{1}{n} \sum x_i t_i.$$

Теперь применим наши знания математического анализа к задаче нахождения минимума функции

$$\min_{(c_0; c_1)} \left\{ \sum_{i=1}^{n} (x_i - c_0 - c_1 t_i)^2 \right\}.$$

Найдем на первом шаге минимум функции, зависящей от одной переменной w, считая c_1 фиксированным параметром (что касается t_i и x_i, то они обозначают фиксированные числа — элементы выборки)

$$f(w) = \sum (x_i - w - c_1 t_i)^2.$$

Найдем точку w, в которой обращается в нуль $f'(w)$. Производная от суммы функций равна сумме производных слагаемых, поэтому

$$f'(w) = \sum \left((x_i - w - c_1 t_i)^2\right)' = \sum 2 \cdot (x_i - w - c_1 t_i) \cdot (-1).$$

Приравняем правую часть к нулю, опустив множители и раскрыв скобки под знаком суммы:

$$\sum (x_i - w - c_1 t_i) = \sum x_i - \sum w - \sum t_i = \sum x_i - nw = 0$$

(мы использовали равенство $\sum w = w + \dots + w = nw$). Решение уравнения $\sum x_i - nw = 0$:

$$w = \frac{1}{n} \sum x_i = \bar{x}.$$

Мы нашли, что *при любом данном угловом коэффициенте* наилучшая из *имеющих данный угловой коэффициент* прямая имеет свободный член $\bar{x}$, т.е. проходит через точку $(0, \bar{x})$ на плоскости[1].

Теперь, вдохновленные успехом, будем искать наилучший угловой наклон для прямых, имеющих этот наилучший свободный член, т.е. найдем минимум зависящей от переменной v функции

$$g(v) = \sum (x_i - \bar{x} - v t_i)^2.$$

Найдем производную по v:

$$g'(v) = \sum (2 \cdot (x_i - \bar{x} - v t_i) \cdot (-t_i)) =$$

$$= -2 \sum (x_i t_i - \bar{x} t_i - v t_i^2) = -2\left(\sum x_i t_i - \bar{x} \sum t_i - v \sum t_i^2\right) =$$

$$= -2\left(\sum x_i t_i - v \sum t_i^2\right).$$

Опустив множители, приравняем правую часть к нулю:

$$\sum x_i t_i - v \sum t_i^2 = 0,$$

[1] Это центр тяжести системы точек (t_i, x_i). В старых переменных он имеет координаты $(\bar{t}, \bar{x})$.

откуда находим

$$v = \frac{\sum x_i t_i}{\sum t_i^2} = \frac{\frac{1}{n}\sum x_i t_i}{\frac{1}{n}\sum t_i^2}.$$

Замечаем, что в числителе стоит выборочная ковариация, а в знаменателе (смещенная) выборочная оценка дисперсии t_i, поэтому окончательно

$$v = \frac{R_{tx}}{S_t^2}.$$

Таким образом,

$$\min_{(c_0, c_1)} \left\{ \sum_{i=1}^{n} (x_i - c_0 - c_1 t_i)^2 \right\} = \sum_{i=1}^{n} (x_i - \hat{c}_0 - \hat{c}_1 t_i)^2 = \Delta^2,$$

где

$$\hat{c}_0 = \bar{x}, \quad \hat{c}_1 = \frac{R_{tx}}{S_t^2}.$$

Минимальное значение, которое мы обозначили символом Δ^2, называется *кажущейся ошибкой* МНК.

Упражнение 8.4. Показать, что в исходных координатах оценки МНК задаются формулами

$$\hat{c}_1 = \frac{R_{tx}}{S_t^2}, \quad \hat{c}_0 = \bar{x} - \hat{c}_1 \bar{t}.$$

8.3.1. Математическая модель

В линейном регрессионном анализе в качестве *математической модели зависимости пары переменных* (t, x) *рассматривается линейная зависимость* $x = c_0 + c_1 t$ *со случайной нормальной ошибкой*, а именно:

$$x_i = c_0 + c_1 t_i + \sigma \xi_i, \quad \xi_i \sim N(0,1), \quad i = 1, 2, \ldots, n, \quad \sum_{i=1}^{n} t_i = 0,$$

где c_0, c_1 и $\sigma > 0$ — фиксированные неизвестные параметры, (t_i, x_i) — пары измерений, а случайные (неизвестные) величины ξ_i, $i = 1, 2, \ldots, n$ имеют стандартное нормальное распределение и независимы.

Допустим, некая реальная ситуация верно отражается данной моделью. Если мы, проведя несколько независимых испытаний, получим n пар результатов (t_i, x_i) и затем найдем точечные оценки параметров

$\hat{c}_0$, $\hat{c}_1$ и $\hat{\sigma}$, то результат не совпадет с реальными параметрами c_0, c_1 и σ, а будет случаным образом отклоняться от них.

Распределение отклонений можно описать и, пользуясь этим описанием, построить доверительные интервалы. Наиболее часто в прикладных исследованиях используется доверительный интервал для c_1: если 0 не входит в доверительный интервал, то угловой коэффициент регрессионной прямой значимо отличается от нуля, а это интерпретируется как наличие реальной систематической функциональной зависимости между t и x.

8.3.2. Доверительные интервалы параметров c_0, c_1 и σ

Точечная оценка $\hat{\sigma}$ среднеквадратичного отклонения σ определяется с помощью кажущейся ошибки Δ^2:

$$\hat{\sigma} = \sqrt{\frac{\Delta^2}{n-2}}, \quad \text{где} \quad \Delta^2 = \sum_{i=1}^{n} (x_i - \hat{c}_0 - \hat{c}_1 t_i)^2.$$

Можно показать, что из определения математической модели параметрической линейной регрессии вытекают следующие утверждения, задающие правила построения доверительных интервалов.

Пусть $\chi^+_\alpha(n-2)$ и $\chi^-_\alpha(n-2)$ - верхний и нижний двусторонние квантили распределения хи-квадрат с $n-2$ степенями свободы. Тогда $P\left(\sigma^-_\alpha < \sigma < \sigma^+_\alpha\right) = 1 - \alpha$, где левый σ^-_α и правый σ^+_α концы доверительного интервала для параметра σ вычисляются по формулам

$$\sigma^-_\alpha = \sqrt{\frac{\Delta^2}{\chi^+_\alpha(n-2)}}, \quad \sigma^+_\alpha = \sqrt{\frac{\Delta^2}{\chi^-_\alpha(n-2)}}.$$

Пусть $t_\alpha(n-2) > 0$ обозначает верхний двусторонний квантиль распределения Стьюдента с $n-2$ степенями свободы для уровня значимости α. Тогда

$$P\left(|c_0 - \hat{c}_0| < \epsilon^0_\alpha\right) = 1 - \alpha, \qquad P\left(|c_1 - \hat{c}_1| < \epsilon^1_\alpha\right) = 1 - \alpha,$$

где радиусы доверительных интервалов

$$\epsilon^0_\alpha = \frac{\hat{\sigma} \cdot t_\alpha(n-2)}{\sqrt{n}}, \qquad \epsilon^1_\alpha = \frac{\hat{\sigma} \cdot t_\alpha(n-2)}{\sqrt{nS^2_t}}.$$

Следовательно, при коэффициенте доверия $1 - \alpha$ концы доверительных интервалов для параметров c_0 и c_1 вычисляются по однотипным формулам

$$c_0^- = \hat{c}_0 - \epsilon_\alpha^0, \qquad c_0^+ = \hat{c}_0 + \epsilon_\alpha^0,$$
$$c_1^- = \hat{c}_1 - \epsilon_\alpha^1, \qquad c_1^+ = \hat{c}_1 + \epsilon_\alpha^1.$$

Можно показать, что при использовании старых переменных радиусы доверительных интервалов имеют вид:

$$\epsilon_\alpha^0 = \frac{\hat{\sigma} \cdot t_\alpha(n-2)}{\sqrt{n}} \cdot \sqrt{1 + \frac{\bar{t}^2}{S_t^2}}, \qquad \epsilon_\alpha^1 = \frac{\hat{\sigma} \cdot t_\alpha(n-2)}{\sqrt{nS_t^2}}.$$

Замечание 2. Для того же множества точек (t_i, x_i) регрессионная прямая, задающая зависимость t от x, не будет совпадать с построенной ранее. В самом деле, если $t_1, \ldots, t_n$ и $x_1 \ldots x_n$ стандартизованные выборки с выборочным коэффициентом ковариации 0,5, то хотя обе регрессионные прямые, выражающие зависимость t от x и x от t, будут проходить через начало координат, но одна из них будет составлять острый угол (тангенс которого равен 0,5) с осью t, а другая — тот же острый угол, но с осью x. Надо иметь в виду, что регрессионная модель подразумевает нечто вроде функциональной зависимости между переменными, что означает их заведомо различный статус.

Послесловие для студентов-гуманитариев и преподавателей математики

При обучении математике студентов гуманитарных факультетов не раз доводилось замечать, что почти у каждого из них неизбежно возникает (чаще озвученный, реже немой) вопрос: "Откуда это и зачем это нужно?". Разумеется, вопросы подобного рода возникают и у студентов естественно-научных и инженерно-технических направлений и специальностей, но именно у будущих студентов-гуманитариев они принимают особенно острые формы; может быть потому, что их отношение к математике особое. Поэтому во времени, отпущенном учебным планом на математическую составляющую, непременно нужно выделить долю для соответствующих пояснений. Мотивация является естественной составной частью содержания преподаваемой дисциплины, хотя и следует признать, что поиск убедительных мотивировок нередко трудная задача.

Вот несколько возможных направлений отыскания нужных мотиваций.

Можно сослаться на признанные авторитеты — ломоносовское "математику за то учить надобно, что она ум в порядок приводит" или на кантово "во всяком специальном учении о природе можно найти лишь столько собственно науки, сколько в нём можно найти математики". Однако здесь нужно иметь в виду, что представление классиков о том, что такое математика, скорее всего разительно отличается от представлений студентов о ней, вынесенных, как правило, из средней школы. А вот раскрыть, о какой математике говорят Ломоносов и Кант, далеко не просто. К тому же ссылки на авторитеты (тем более из XVIII века) вряд ли покажутся в нашем веке студентам достаточно убедительными.

Так сложилось, что чаще всего изложение совокупности сведений из математики предлагается студентам в виде повествования, в котором даются ответы на непоставленные вопросы. И возникающее недоумение студентов можно понять. Когда-то это были ответы на животрепещущие вопросы, которые позже сменили новые актуальные вопросы. Полученные на них ответы привели к новым вопросам, и так продолжа-

лось довольно долго. Затем из ответов была соткана логически прочная ткань, в которой вопросам не было места, ибо всем своим неформальным видом они выбивались из стройных повествовательных цепочек. Процесс накапливания математических познаний шёл совсем не так, как это принято излагать на лекциях и семинарских занятиях. Наши великие предшественники искали ответы на волновавшие их вопросы и одновременно учились правильно ставить сами эти вопросы. Восстанавливая, хотя бы и кратко, этот естественный процесс размышлений и тем самым вовлекая в этот процесс студентов, можно надеяться на проявление определённого интереса к затрагиваемым вопросам и, как следствие, на их понимание.

Поступившему в высшее учебное заведение предстоит несколько лет овладевать основами знаний и иными премудростями, чтобы на излёте обучения получить более или менее полные сведения о том, что же представляет собой их будущая специальность. А вот в самом процессе обучения большинству студентов далеко не всегда достаточно ясно, чем именно вызвано разнообразие изучаемых предметов и каким именно образом они способны взаимно дополнять и обогащать друг друга. Поэтому формирование мотивации, опирающейся на интерес студента к выбранной специальности или выбранному направлению, по отношению к математической составляющей проходит сложно — дело в том, что изучение элементов математики чаще всего отнесено на младшие курсы, когда будущий специалист имеет о своей специальности весьма приблизительное представление, да и в самом процессе обучения ещё не дошёл до содержательных задач, где умело применённый математический инструментарий способен внести свою неповторимую лепту. Обучаясь математическим приёмам, студент не готов к восприятию содержательных профессиональных задач, и потому показать действенность и возможности предлагаемых ему математических подходов к их разрешению нередко просто невозможно.

В книге предпринята определённая попытка применить эти подходы к мотивировке отобранного материала. Подход, основанный на связи с будущей специальностью, проявляется в том, что авторы видят среди своих читателей в первую очередь студентов психологических факультетов и специальностей. Вместе с тем описание естественности вводимых математических понятий и связей между ними столь узкой привязки уже не имеет, так как здесь авторы делают попытку показать, как возникали те или иные вопросы и как искались и находились ответы на них.

Конечно, можно было найти и другие примеры, которые бы звучали столь же и даже более убедительно. И если преподающий математику отнесётся к решению этой задачи с любовью и уважением к студентам, то его успех обеспечен.

Е.В. Шикин

Приложение

Статистические таблицы

Таблица А. Распределение Стьюдента
Доверительные границы для t с f степенями свободы

f	Двухсторонние границы				
	0.1	0,05	0,02	0,01	0,001
1	6,314	12,710	31,820	63,660	636,600
2	2,920	4,303	6.965	9,925	31,600
3	2,353	3,182	4,541	5,841	12,920
4	2,132	2,776	3,747	4.604	8,610
5	2,015	2,571	3,365	4,032	6,869
6	1,943	2,447	3,143	3,707	5,969
7	1,895	2,365	2,998	3,499	5,408
8	1,860	2,306	2,896	3.355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2.624	2,977	4.140
15	1,753	2,131	2,602	2,947	4,073
16	1,746	2,120	2,583	2.921	4,015
17	1,740	2.110	2,567	2,898	3,965
18	1,734	2,101	2,552	2,878	3,992
19	1,729	2,093	2,539	2,861	3,883
20	1,725	2,086	2,528	2,845	3,850
21	1,721	2,080	2.518	2,831	3,819
22	1,717	2.074	2,508	2,819	3,792
23	1,714	2,069	2,500	2,807	3,767
24	1,711	2,064	2,492	2,797	3,745
25	1,708	2.060	2.485	2,787	3,725
26	1,706	2,056	2,479	2,779	3,707
27	1,703	2,052	2,473	2,771	3,690
28	1,701	2,048	2,467	2.763	3,674
29	1,699	2,045	2,462	2,756	3,659
30	1,697	2,042	2,457	2.750	3,646
40	1,684	2,021	2,423	2,704	3,551
50	1,676	2,009	2.403	2,678	3,495
60	1,671	2.000	2,390	2,660	3,460
80	1,665	1,990	2.374	2,639	3,415
100	1,661	1.984	2,365	2,626	3,389
∞	1,645	1.960	2,326	2,576	3,291
	0,05	0.025	0.01	0,005	0,0005
	Односторонние границы				

Таблица Б. Распределение χ^2

Доверительные границы для t с f степенями свободы

	0,2	0,1	0,05	0,01
1	1,642	2,706	3,841	6,635
2	3,219	4,605	5,991	9,210
3	4,642	6,251	7,815	11,345
4	5,989	7,779	9,488	13,277
5	7,289	9,236	11,070	15,086
6	8,558	10,645	12,592	16,812
7	9,803	12,017	14,067	18,475
8	11,030	13,362	15,507	20,090
9	12,242	14,684	16,919	21,666
10	13,442	15,987	18,307	23,209
11	14,631	17,275	19,675	24,725
12	15,812	18,549	21,026	26,217
13	16,985	19,812	22,362	27,688
14	18,151	21,064	23,685	29,141
15	19,311	22,307	24,996	30,578
16	20,465	23,542	26,296	32,000
17	21,615	24,769	27,587	33,409
18	22,760	25,989	28,869	34,805
19	23,900	27,204	30,144	36,191
20	25,038	28,412	31,410	37,566
21	26,171	29,615	32,671	38,932
22	27,301	30,813	33,924	40,289
23	28,429	32,007	35,172	41,638
24	29,553	33,196	36,415	42,980
25	30,675	34,382	37,652	44,314
26	31,795	35,563	38,885	45,642
27	32,912	36,741	40,113	46,963
28	34,027	37,916	41,337	48,278
29	35,139	39,087	42,557	49,588
30	36,250	40,256	43,773	50,892

Таблица B. Распределение Вилкоксона

Нижние граничные значения

n	Односторонние 0,05 Двухсторонние 0,1	Односторонние 0,025 Двухсторонние 0,05	Односторонние 0,01 Двухсторонние 0,02
5	1		
6	2	1	
7	4	2	0
8	6	4	2
9	8	6	3
10	11	8	5
11	14	11	7
12	17	14	10
13	21	17	13
14	26	21	16
15	30	25	20
16	36	30	24
17	41	35	28
18	47	40	33
19	54	46	38
20	60	52	43
21	68	59	49
22	75	66	56
23	83	73	62
24	92	81	69
25	101	90	77
26	110	98	85
27	120	107	93
28	130	117	102
29	141	127	111
30	152	137	120

Таблица Г. Распределение Манна—Уитни

Нижние границы (начало)

Односторонние при $\alpha = 0,05$
Двухсторонние при $\alpha = 0,1$

n_1 \ n_2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																	0	0
2					0	0	0	1	1	1	1	2	2	3	3	3	4	4
3	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
4		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
5			4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
6				7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
7					11	13	15	17	19	21	24	26	28	30	33	35	37	39
8						15	18	20	23	26	28	31	33	36	39	41	44	47
9							21	24	27	30	33	36	39	42	45	48	51	54
10								27	31	34	37	41	44	48	51	55	58	62
11									34	38	42	46	50	54	57	61	65	69
12										42	47	51	55	60	64	68	72	77
13											51	56	61	65	70	75	80	84
14												61	66	71	77	82	87	92
15													72	77	83	88	94	100
16														83	89	95	101	107
17															96	102	109	115
18																109	116	123
19																	123	130
20																		138

Таблица Г. Распределение Манна—Уитни

Нижние границы (окончание)

Односторонние при $\alpha = 0{,}01$
Двухсторонние при $\alpha = 0{,}02$

n_1	n_2															
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																
2									0	0	0	0	0	0	1	1
3			0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6		3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7			6	8	9	11	12	14	16	17	19	21	23	24	26	28
8				10	11	13	15	17	20	22	24	26	28	30	32	34
9					14	16	18	21	23	26	28	31	33	36	38	40
10						19	22	24	27	30	33	36	38	41	44	47
11							25	28	31	34	37	41	44	47	50	53
12								31	35	38	42	46	49	53	56	60
13									39	43	47	51	55	59	63	67
14										47	51	56	60	65	69	73
15											56	61	66	70	75	80
16												66	71	76	82	87
17													77	82	88	93
18														88	94	100
19															101	107
20																114

Таблица Д. Значения функции $\Phi(x)$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,1	0,040	1,1	0,364	2,1	0,482
0,2	0,079	1,2	0,385	2,2	0,486
0,3	0,118	1,3	0,403	2,3	0,489
0,4	0,155	1,4	0,419	2,4	0,492
0,5	0,192	1,5	0,433	2,5	0,494
0,6	0,226	1,6	0,445	2,6	0,495
0,7	0,258	1,7	0,455	2,7	0,497
0,8	0,288	1,8	0,464	2,8	0,497
0,9	0,316	1,9	0,471	2,9	0,498
1,0	0,341	2,0	0,477	3,0	0,499

Учебное издание

Анатолий Николаевич Кричевец
Евгений Викторович Шикин
Аркадий Георгиевич Дьячков

Математика для психологов

Учебник

Подписано в печать 28.02.2003. Формат 60x88/16. Печать офсетная
Усл. печ. л. 23,0. Уч.-изд. л. 19,8. Тираж 10000 экз. Изд. № 658. Заказ 1347

ИД № 04826 от 24.05.2001 г.
ООО «Флинта», 117342, г. Москва, ул. Бутлерова, д. 17-Б, комн. 332
Тел/факс (095) 334-82-65, тел. (095) 336-03-11
E-mail: flinta@mail.ru
WebSite: www.flinta.ru

ЛР № 020297 от 23.06.1997 г.
Издательство «Наука», 117997, ГСП-7, Москва В-485
ул. Профсоюзная, д. 90

Отпечатано с готовых диапозитивов во ФГУП ИПК
«Ульяновский Дом печати». 432980, г. Ульяновск, ул. Гончарова, 14

известностью и популярностью у студентов и преподавателей вузов России и стран СНГ. Студенты любят нашу литературу, она является хорошим помощником как в учебном процесса, так и при подготовке к экзаменам.

●

Издаваемая литература выходит в сериях: «Библиотека педагога-практика», «Библиотека социального работника», «Библиотека социального педагога», «Библиотека школьного психолога», «Преподавание психологии в школе», «Библиотека психолога», «Библиотека логопеда», «Библиотека студента» «Библиотека юриста», «Библиотека экономиста», «Библиотека менеджера» и др. С 1995 г. издается **уникальная серия «Психологи Отечества»** — избранные психологические труды выдающихся отечественных ученых-психологов XIX—XX веков (в 70 томах), не имеющая аналогов в мире.

А.Н. Кричевец, Е.В. Шикин, А.Г. Дьячков

Математика для психологов

Учебник

Кричевец Анатолий Николаевич — к.ф.-м.н., доктор философских наук, старший научный сотрудник лаборатории нейропсихологии факультета психологии МГУ им. М.В. Ломоносова

Шикин Евгений Викторович — д.ф.-м.н., профессор, заведующий кафедрой математических методов в управлении факультета Государственного управления МГУ им. М.В. Ломоносова

Дьячков Аркадий Георгиевич — д.ф.-м.н., профессор кафедры теории вероятностей механико-математического факультета МГУ им. М.В. Ломоносова

ISBN 5-89349-400-8

9785893494006

БИБЛИОТЕКА СТУДЕНТА

Математика для психологов